UNDERSTANDING CYCLING BEHAVIOUR THROUGH VISUAL ANALYSIS OF A LARGE-SCALE OBSERVATIONAL DATASET

Roger Beecham
giCentre, Department of Computer Science
City University London
roger.beecham.1@city.ac.uk

A thesis submitted for the degree of Doctor of Philosophy in Geographic Information Science

July 2014

Abstract

The emergence of third-generation, technology-based public bikeshare schemes offers new opportunities for researching cycling behaviour. In this study, data from one such scheme, the London Cycle Hire Scheme (LCHS), are analysed. Algorithms are developed for summarising and labelling cyclists' usage behaviours and tailored visual analysis applications are designed for exploring their spatiotemporal context.

Many of the research findings provide support to existing literature, particularly around gendered cycling behaviour. As well as making more discretionary journeys, women appear to preferentially select parts of London associated with greater levels of safety; and this is true even after controlling for geodemographic differences and levels of LCHS cycling experience. One hypothesis is that these differences represent diverging attitudes and perceptions. After developing a technique for identifying cyclists' workplaces, these differences might also be explained by *where* cyclists need to travel for work and other facilities. An additional explanation is later offered that relates to the *nature* of cyclists' estimated routes. The size and precision of the LCHS dataset allows under-explored aspects of behaviour to be investigated. Group cycling events – instances where two or more cyclists make journeys together in space and time – are labelled and analysed on a large scale. For certain types of cyclist, group cycling appears to encourage more extensive spatiotemporal cycling behaviour and there is some evidence to suggest that group cycling may help initiate scheme usage.

The domain-specific findings, emerging research questions and also behavioural classifications are this study's principal and unique contribution. A second contribution relates to the analysis approach. This is a data-driven study that takes a large dataset, measuring use of a relatively new cycle facility, and uses it to engage with research questions that are typically answered with very different datasets. There is some uncertainty around how discriminating and generalisable LCHS cycle behaviours may be and which variables,

either directly measured or derived, might delineate those behaviours. Visual analysis techniques are shown to be effective in this more speculative research context: numerous behaviours are very quickly explored and understood. These techniques also enable a set of colleagues with relatively limited analysis experience, but substantial domain knowledge, to participate in the analysis and a general argument is made for their use in other, interdisciplinary analysis contexts.

Acknowledgments

Despite reading various harrowing accounts of how to 'survive' the PhD process, I've found the last three years of study to be incredibly enjoyable and stimulating. I am grateful to a number of colleagues and friends for their support and inspiration throughout this period.

First and foremost, I'd like to thank my supervisor, Jo Wood. As well as his time, expertise, ideas and critique, Jo's practical contributions have been instrumental. Without his input, the successful working relationship with Transport for London (TfL) wouldn't have happened and the research outputs certainly wouldn't have been presented so widely. Jo's commitment to high quality research and teaching is truly inspiring. It has been a real privilege to work with Jo and after spending some time 'evaluating my career options', it is through him that I have developed a renewed and growing interest in research.

I'd like to give general thanks to all those at the giCentre and other occupants of A304. Thanks in particular to Aidan Slingsby, Iain Dillingham and Alex Kachkaev, to whom I regularly directed various somewhat ill-defined questions. Often these led to constructive and useful discussions; if I'm honest, they were also successful diversion tactics. Thanks also to Sarah Goodwin and Ali Ramathan for similar reasons. Special thanks should go to Jason Dykes for his encouragement and interest throughout the research project, but also in helping with my initial research proposal. I am grateful to City University for funding the PhD through a university Studentship, providing an excellent working environment and supporting my conference attendance.

I'd like to thank colleagues at TfL not only for enabling full access to the bikeshare data, but for their interest, incredibly valuable interpretation and insight. In particular, I'd like to thank James Hiett for his work in securing the data sharing agreement and Audrey Bowerman and Sarah Burr for their enthusiasm and policy-related expertise.

I am ever grateful to my mum and dad for their uncompromising support and advice

and hopefully their ‘60s grammar school education when it came to proof-reading this document. And last but not least, my fiancée Sam, for sitting through at least five presentation rehearsals on this work (see *Invited talks* section below).

Publications

The materials, ideas and graphics in this document have appeared previously in the publications and talks listed below.

Journal articles

- Wood, J., **Beecham, R.** & Dykes, J. (in press) Moving beyond sequential design: Reflections on a rich multi-channel approach to data visualization. *IEEE Transactions on Visualization and Computer Graphics*.

- **Beecham, R.** & Wood, J. (2014) Characterising group-cycling journeys using interactive graphics. *Transportation Research Part C: Emerging Technologies*, 47(October), pp.194-206. doi: 10.1016/j.trc.2014.03.007.

- **Beecham, R.** & Wood, J. & Bowerman, A. (2014) Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, 47(September), pp.5-15. doi: 10.1016/j.compenvurbsys.2013.10.007.

- **Beecham, R.** & Wood, J. (2014) Exploring gendered cycling behaviours within a large-scale behavioural dataset. *Transportation Planning and Technology*, 37(1), pp.83-97. doi: 10.1080/03081060.2013.844903.

Book chapters

- **Beecham, R.** (in preparation) Using bikeshare datasets to improve urban cycling experience and research urban cycling behaviour. In Gerike, R., Cox, P., de Geus, B. & Parkin, J. (in preparation) *The future of cycling*, Ashgate, London, UK.

Conference papers

- **Beecham, R.** & Wood, J. (2014) Towards confirmation? Deriving and analysing routing information from an origin-destination bikeshare dataset. 46th Annual Universities Transport Study Group Conference, 6 - 8 January 2014, Newcastle, UK.

Invited talks

- **Beecham, R.** & Wood, J. (2014) Discovering bikeshare cycle behaviours through interactive visual analysis. Or why pictures are a necessary part of big data analysis. London School of Hygiene and Tropical Medicine Transport & Health Group Seminar, 14 January 2014, London, UK.
- **Beecham, R.** (2013) Exploratory visualization for discovering data stories. Hacks versus Hackers Meetup, 27 November 2013, London, UK.
- **Beecham, R.** (2013) Data visualization. The Power of Data, PPA Digital Publishing Conference 2013, 18 September 2013, London, UK.
- **Beecham, R.** (2013) Visualization for better data analysis. Transport data visualisations, Transport Statistics User Group, 20 March 2013, London, UK.
- **Beecham, R.** (2013) Exploring gender and cycle behaviour in a large-scale dataset. Urban Digital: GIS Mapping and Technology, King's College London, 15 February 2013, London, UK.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Abstract

The emergence of third-generation, information-technology based bikeshare schemes offers a new means of researching observed cycling behaviour on a large scale. A complete set of cyclist-level usage data from one such scheme, the London Cycle Hire Scheme (LCHS), have been made available by Transport for London (TfL) for use in this research. Analysing these data, distinct customer cycling behaviours are identified, classification techniques for labelling those behaviours are developed and possible motivations behind observed behaviours are suggested. The principal research question asks: How, and to what extent, can the LCHS datasets be used to contribute to current research on cycling behaviour in Transport Studies? A detailed set of domain-specific research findings, emerging research questions and also behavioural classifications are the main and unique contribution. In researching individual cyclists' behaviour, the LCHS dataset is used for a purpose that may be different from that for which it was initially intended. Since *individual* bikeshare cycling behaviours have yet to be studied in detail, there is some uncertainty around how usage might be structured and which variables, either directly measured or derived, might help delineate those behaviours. A second contribution is an argument for, and demonstration of, the use of visual techniques in this more speculative research context.

1.1 Research context

The many health-related, economic and time-saving benefits of cycling have brought about a cycling renaissance (Pucher & Buehler 2012) and greater political ambition to increase cycling across society (APPC 2013). A challenge for the United Kingdom (UK), as in many car-oriented countries, is that levels of cycling as a whole remain persistently low (Department for Transport 2013), with the recent uptake not evenly distributed geographically or socially (Goodman 2013). This fact has precipitated a growing academic interest in understanding cycling behaviour and especially for policy implementation, an interest in the factors that motivate and discourage cycling within cities (Pucher & Buehler 2012).

Much of this existing research has been led by the social sciences and has relied on traditional, actively collected datasets (Pooley et al. 2011, Aldred 2012). Using data from large social surveys, a number of studies have successfully identified varying public attitudes towards cycling held by different sections of society (Davies et al. 2001, Gatersleben & Appleton 2007), whilst more detailed qualitative and ethnographic approaches have attended to the very complex social, economic, cultural and physiological circumstances that affect individual behaviour (Davies et al. 1997, Garrard et al. 2008, Aldred 2012). Research into observed, rather than claimed or stated, cycling behaviour has been more limited both in size and ambition. Automatic Cycle Traffic Counters (ACTCs) provide one means of studying 'actual' behaviour. Historical data collected from such counters have allowed relatively concrete conclusions to be made about the impact of seasonal and weather variations on aggregate levels of cycling (Thomas et al. 2013, Tin et al. 2012, Niemeier 1996). A number of Global Positioning System (GPS)-based studies (Dill & Gliebe 2008, Broach et al. 2012) have also considered more detailed aspects of observed spatial travel behaviour.

There are limitations to the observational data collected from ACTCs and also the GPS-based studies. With ACTCs, cycle behaviours are only monitored at a given spatial point, most counters are not able to record direction (Gordon 2012) and clearly nothing is known about the trajectory of journeys or, perhaps more importantly, the people making them. Whilst Global Positioning Systems (GPS) technology does offer this very detailed information, due to the cost, complexity and logistical challenges of producing these data, such work is generally associated with datasets that are small in scale. Problems of self-selection bias, where particular types of cyclist volunteer to participate in these

studies (Dill & Gliebe 2008) and social-desirability bias, where research participants may alter their behaviours because they know they are being monitored (Dill 2006), are genuine concerns.

The recent growth in so-called *third generation* (Shaheen et al. 2012), information-technology based urban bikeshare schemes offers a new means of researching observed cycling behaviour on a large scale. In many recent schemes, data on usage are continually reported to central databases. Researchers working within data mining (Froehlich et al. 2008, Jensen et al. 2010, Kaltenbrunner et al. 2010, Borgnat et al. 2011, Lathia et al. 2012, Côme & Oukhellou in press), information visualization (Wood et al. 2011), geography (O'Brien et al. 2014, Goodman & Cheshire in press) and public health (Fuller et al. 2012, Ogilvie & Goodman 2012, Woodcock et al. 2014) have queried these data to identify detailed patterns of scheme use over space and time. The scope and nature of the early bikeshare analysis has nevertheless been constrained by the level of detailed information made publicly available. In many cases data were harvested from the web, where local transport authorities publish in real-time the number of available bikes at individual bikeshare docking stations (Froehlich et al. 2008, Lathia et al. 2012). Others gained access to journey records, including journey origin-destination (OD) and start and end times (Jensen et al. 2010, Borgnat et al. 2011, Wood et al. 2011, O'Brien et al. 2014). With the exception of Ogilvie & Goodman (2012), Woodcock et al. (2014) and Goodman & Cheshire (in press), researchers did not have access to a customer database and bikeshare journeys could not be linked back to individual customers. Individual cyclists' journey histories could therefore not be identified, limiting the extent to which these bikeshare data could be used to engage with the more complex questions around motivations and barriers to cycling that have been studied using more traditional, actively-collected datasets.

Working with policy makers at *Transport for London* (TfL)[1], these more detailed data on usage of the London Cycle Hire Scheme (LCHS) have been made available for specific use in this research. Customer records reporting a unique customer identifier, gender and full postcode customers registered with, have been made available. So too has a complete set of user journeys. Here, an OD pair with associated timestamps is recorded for every journey that is made. The customers and journeys datasets can be related directly with a customer identifier variable, which means that *individual* cyclists' behaviours can be identified for the entire registered customer population. This amounts to 13.9 million

[1] `http://www.tfl.gov.uk/`

journeys made by almost 150,000 registered customers between 30th July 2010 and 27th April 2013: a record of observed, individual cycling behaviour of unprecedented size.

1.2 Analysis objectives

An overriding aim is to investigate how the LCHS data can be used to contribute to, and extend, existing research into cycling behaviour:

- **Global objective**: To develop a set of research findings and emerging research questions that contribute to, and extend, existing research into cycling behaviour in Transport Studies.

As a timed origin-destination dataset, many possibilities exist for using the LCHS data to research various aspects of behaviour on a large scale. A substantial aspect of this research is a data analysis of the LCHS datasets to characterise in detail how cyclists use the scheme. Through extensive spatiotemporal querying, different types of cycling behaviours are identified and subsequently labelled and, studying the context under which these journeys are made, inferences about possible journey purpose and the motivations behind observed behaviours are suggested. Three research objectives guide this data analysis:

- **Objective 1**: To identify distinct customer cycling behaviours through exploring space-time patterns of travel.
- **Objective 2**: To develop classification techniques for labelling behaviours identified through exploratory analysis.
- **Objective 3**: To suggest and investigate possible explanations for observed behaviours.

1.3 Analysis approach

In contrast to much of the literature on urban cycling behaviour (see section 1.1), the LCHS dataset is used for a purpose that may be different from that for which it was

initially intended. The LCHS usage datasets are not maintained solely for the purpose of social research – for understanding individual-level cycling behaviour. Rather than first specifying research questions and carefully structuring empirical data collection so that these stated questions can be answered directly, initial insights and ideas about behaviour must be derived from exploring already existing travel records. For example, this research aims to *infer* important aspects of cycle behaviour, such as journey purpose, by attending to individuals' historical scheme usage and the space-time context under which journeys are made. By contrast, in traditional, actively collected travel surveys, such information as journey purpose, and deeper motivations for cycling, might be stated explicitly. Also, unlike most survey-based research, this project considers a large, but very particular population of cyclists with a relatively limited set of background information on those individuals. When attempting to explain observed behaviours, it is therefore important to consider which demographic and other variables are directly measured, which might be accessed through leveraging external datasets and whether further contextual information might be *derived* through studying individual-level scheme usage.

Throughout, visual analysis techniques are shown to be highly effective in this more speculative research context. By visually exploring the LCHS data at various spatial and temporal resolutions and by pre-computed behavioural variables, numerous usage patterns are quickly explored and delineated. On the back of this exploratory analysis, more specific research questions and hypotheses are explored visually and tested quantitatively. Chapter 2 elaborates on the visual analysis approach, Chapters 3 and 4 discuss how visual analysis techniques are used for exploratory data analysis and Chapter 5 demonstrates how tailored interactive visual interfaces are used to support analytical decision-making.

1.4 Research contributions

This is one of the first extensive studies of its kind: one of the first to use data from a bikeshare scheme to study *individual* spatiotemporal cycle behaviours in detail. In Chapter 4, large-scale empirical evidence is provided to support well-rehearsed arguments around gender and urban cycle behaviour, along with two new insights on this theme. Chapter 5 discusses new techniques created for labelling commuting travel behaviours in the LCHS dataset and which might be applied to other large, passively collected OD datasets. Chapter 6 contributes empirical evidence to a research theme that has yet

to be studied within the existing literature, but was briefly proposed by O'Brien et al. (2014) in their data analysis of bikeshare schemes across cities: that of 'group cycling'. Finally, Chapter 7 demonstrates how early exploratory findings associated with spatial travel behaviours might be studied in more detail and discusses the limits to making explanatory claims about observed LCHS cycle behaviours. The description of analysis techniques and research findings are therefore this study's main contribution:

- **Primary contribution**: A uniquely large-scale, detailed and spatiotemporal analysis of observed urban cycle behaviours that contributes a novel set of research findings.

A secondary contribution relates to the visual analysis approach. By visually representing the LCHS datasets, spatiotemporal structures of behaviour are explored and identified. In Chapter 5, visual analytics software is also used in more involved analysis activities: in selecting appropriate spatial analysis techniques for labelling commuting behaviours and making informed decisions about the thresholds and parameters used in those techniques. This analysis might have been completed using non-visual means. The benefit of instead developing tailored visualization software is demonstrated in two ways. Firstly, in enabling a data-driven analysis process. Spatiotemporal structures of behaviour are quickly identified, new themes of analysis are queried 'on the fly' and empirically-informed analysis decisions are made using the visual analysis applications that appear in Chapters 3 and 5. The second has implications for others working particularly in data-driven social research: the visual analysis software supports a process of collective data analysis with a group of colleagues at TfL who have substantial subject knowledge, but relatively limited analytical experience.

- **Secondary contribution**: An argument for, and demonstration of, the use of interactive visual analysis techniques in an applied, data-driven research setting.

1.5 Research questions

These contributions relate both to research findings and techniques and this study might be located within the so-called 'computational social sciences' (Lazer et al. 2009, King 2011). A large, behavioural dataset is processed and used to analyse an existing research problem within an established social science domain. A criticism levelled at data-driven

social science studies is that too often they privilege the application and development of computing techniques over genuine research questions. As a result, they risk only affirming existing social theories rather than generating new insights into those theories (Giles 2012, Miller 2010, Watts 2013). This research perhaps overcomes these problems by locating data analysis themes within relevant literature, closely involving domain specialists (TfL) in the analysis and contributing new insights to current research questions in Transport Studies. That research findings have already been published in high profile academic journals in Transport Studies – *Transportation Research Part C, Transportation Planning and Technology* – as well as Geography – *Computers, Environment and Urban Systems* – and Information Visualization – *Transactions on Visualization and Computer Graphics* – is further evidence of the substantive contribution of this data analysis study. A high-level question implicit throughout this research is:

- **Overall research question**: How, and to what extent, can the LCHS dataset be used to contribute to current research on cycling behaviour in Transport Studies?

Three more specific research questions (RQs) are used to structure and organise analysis activities and that relate to the objectives listed in Section 8.1:

- **RQ1**: How are individual LCHS customers' usage behaviours differentiated?
- **RQ2**: How might usage behaviours be labelled?
- **RQ3**: To what extent can identified behaviours be explained?

1.6 Thesis outline

After elaborating on visual approaches to analysis and a detailed description of datasets and exploratory data analysis techniques, this document is organised around a data analysis of the LCHS data. Each analysis chapter is thematically distinct, with a discrete set of research questions. However, as the enquiry progresses, research findings, and data-driven hypotheses for explaining observed behaviours, become more detailed: chapters move from a speculative, exploratory analysis of research themes, to a more detailed study of formally labelled behaviours.

In **Chapter 2**, some of the challenges, already discussed in Section 1.3, associated with using the LCHS dataset to study cycling behaviour are repeated, along with a critical discussion of visual analysis techniques and a justification for their use in this research.

Chapter 3 describes the LCHS dataset in detail: how data are recorded and also external information used to supplement the customer database. A number of behavioural variables for summarising individual cyclists' scheme usage are created and the main visual analysis application for exploring the LCHS dataset is introduced. Decisions around these behavioural variables, and design decisions for the visual analysis application, are justified with recourse to literature within Transport Studies and Information Visualization.

Chapter 4 is the first substantive findings section. The current literature on gender and urban cycle behaviour is large and the chapter focusses on differences in men's and women's cycle behaviours using the software described in Chapter 3. Substantial differences between men's and women's usage are found that might relate to differences in the types of men and women subscribing to the scheme, but also possibly to more fundamental differences in men's and women's approaches to cycling. The chapter concludes by reflecting on the level and detail of findings that were achieved through simply exploring the LCHS dataset.

Chapter 5: In Chapter 4, claims about apparent commuting behaviour are made by visually scanning spatiotemporal travel behaviours. As the analytical enquiry progresses, observed behaviours are studied more formally. In Chapter 5 a technique for automatically labelling commuting journeys is developed and commuting behaviours are investigated in some detail. The approach to labelling commuting journeys may be applied to other origin-destination transport datasets, but the chapter concludes by focussing on the implications of the workplace classification on this study and on earlier attempts at *explaining* bikeshare cycling behaviours.

The LCHS dataset provides a total, population-level record of customers' scheme usage. **Chapter 6** demonstrates how this fact allows a new, previously under-researched aspect of cycling behaviour to be studied: that of 'group cycling'. Group cycling is defined as journeys made between at least a pair of bikeshare cyclists together in space and time. Once labelled, group-cycling journeys are explored using the software introduced in Chapter 3. A new application is also created here that allows individual group-cycling networks to be explored over space and time. The chapter concludes by reflecting on the

findings and their implications for wider research in Transport Studies.

Chapter 7: Many findings from this analysis relate to interesting and distinct *spatial* cycle behaviours. A limitation here is that with only the origins and destinations of cycle journeys, nothing is known about the nature of likely routes that might be encountered by cyclists. In Chapter 7, an attempt is made to address this issue by collecting information on estimated routes for every cycled OD pair in the dataset. From this information, heuristics on the nature of routed journeys are collected. A problem with this approach is that there is no means of knowing how closely estimated routes relate to customers' actually cycled routes. The chapter therefore focusses on an aspect about which there is greater certainty: the bridge that is suggested by the routing algorithm for journeys that involve a river crossing. More detailed explanations for observed spatial travel behaviours are offered as a result of this analysis, along with a discussion around the extent to which these explanations might be formally described and quantified.

Chapter 8: The Conclusion chapter returns to the three objectives introduced in Section 1.2. The study's domain-specific contributions are outlined, along with an argument for the analysis approach. Some time is spent reflecting on the implications for academic research in Transport Studies, applied Information Visualization and the data-driven social sciences, but also for those working in public policy who wish to promote urban cycling and those in operations responsible for the running of bikeshare schemes. By critically engaging with the limits of this research as well as its contributions, an immediate research agenda is also specified. The chapter concludes by stating the *thesis* argued from this data analysis study.

1.7 Use of literature

This study's primary contribution (Section 1.4) is its empirical findings. The aim is to develop a set of research findings that support, relate to and further existing research within a sub-discipline of Transport Studies that seeks to understand how and why individuals cycle in cities. That the study's findings are sufficiently located within the literature of this domain is therefore an important requirement. An attempt is made to cite relevant domain literature throughout this document and particularly in the analysis chapters. In Chapter 4, for example, relevant studies are used to help identify substantive sub-themes for analysis and validate early research findings. This early validation is

particularly important since, as an entirely new and distinct set of cycle facilities, there is no knowledge of how typical LCHS cycling might be of 'normal', non-bikeshare cycling. Each of the thematic chapters therefore begins with a review of relevant literature from the Transport Studies domain.

A semi-systematic search strategy was used to identify the domain literature that appears in these chapters. Here, an initial list of keywords was created and these search terms entered into three bibliographic databases: *GoogleScholar*[1], *ScienceDirect*[2] and *Mendeley*[3]. As well as an emerging set of keywords, the most frequently returned authors and journals were identified. Decisions were made based on a paper's relevance, quality (its currency, citation count and journal citation indices) and also from browsing the reference lists of papers retrieved from these searches (Rugg & Petre 2007). The same technique was used to identify the early papers that sought to use bikshare cycling behaviour for analysis and which were discussed in Section 1.1. A more detailed exposition of these early studies and a further argument for how the analysis discussed in this document relates to those studies appears in Beecham (in preparation).

1.8 Moving forward

As the outline above suggests, the substantive contribution of this research is a data analysis of LCHS cycling behaviours. As the analysis chapters progress, an increasingly detailed profile of behaviour is provided and throughout, the link between the analysis covered in this research and current literature on cycling in Transport Studies is made explicit. An important aspect of this work, which makes it distinct from much of the existing literature on urban cycling behaviour discussed in Section 1.1, is that an under-explored dataset, measuring use of an entirely new set of cycle facilities, is analysed. In the following two chapters, the challenges associated with using the LCHS dataset to research cycling behaviour are addressed: Chapter 2 considers how visual techniques might offer support in this more uncertain data analysis context and in Chapter 3, some time is spent describing the LCHS dataset and early exploratory analysis design.

[1] `http://scholar.google.com`
[2] `http://www.sciencedirect.com`
[3] `http://www.mendeley.com`

Chapter 2

Visual approach to data analysis

Abstract

Unlike many other studies of cycling behaviour, this research relies on a dataset that was not collected solely for the purpose of doing social research. There is some uncertainty as to the nature and detail of research themes that might be addressed, but also as to the specific information, either already existing, external or derived, that might be used to answer research questions. Visual approaches to analysis are particularly suited to this more speculative analysis context. Designing interactive visual interfaces, patterns of cycling behaviour are quickly discovered and themes within the LCHS dataset explored. New hypotheses about behaviour are suggested as a result of this exploratory analysis, along with a more concrete set of research questions and analysis tasks. The use of visual data analysis methods therefore helps in moving to a point where more specific research questions are addressed and this is reflected in some of the later chapters. Within information visualization, this approach might be regarded as a *design study*: visual analysis software is designed to tackle an applied research problem. A pitfall particular to design studies, and to 'computational social science' research, is that, preoccupied with novelty in visual design, they fail to address problems within a target domain. This issue is perhaps sidestepped in this research by consciously designing and publishing analysis outputs within Transport Studies and closely involving colleagues at TfL, themselves specialists in transport policy, in the analysis.

2.1 Dataset and task uncertainty

This study aims to contribute new insights into research investigating how and why individuals cycle within cities. This literature was briefly introduced in Chapter 1 and more detailed reference is made to specific studies in the themed analysis sections. In most of this previous research, empirical datasets were collected for the purpose of studying specific aspects of cycling behaviour. For example, a researcher might be interested in the association between gender and claimed motivations and barriers to cycling (Heesch et al. 2012). S/he therefore designs a questionnaire and obtains a sample of research participants of appropriate structure and size such that these associations can be analysed within a statistical framework. Any limitations around the scope of analysis and strength of research findings are known in advance; so too, perhaps, are the specific analysis techniques likely to be used in deriving insights from the survey data.

The context under which this data-driven study is completed is necessarily different. The LCHS usage datasets were not necessarily generated for the purpose of doing social research. The customer database has a relatively sparse set of personal attribute information (see Chapter 3) and whilst the record of bikeshare journeys is spatially and temporally precise, important information such as journey purpose is not given. Further demographic information could be added by using external datasets and, through more detailed behavioural analysis, information on journey motivation might be derived. However, in studying LCHS cycling behaviour, this research focusses on a very particular population of cyclists with access to a very particular form of cycling. *Individual* customers' spatiotemporal usage behaviours have yet to be studied in detail within the existing bikeshare research (see Chapter 1). There is some uncertainty around whether a discernible structure might exist within the LCHS and whether the scheme is used sufficiently frequently by individuals for regular patterns of activity, or genuine *travel* behaviours, to be explored.

2.2 *Design study* method

Sedlmair et al. (2012) argue that visual approaches to data analysis are particularly suited to such research contexts: where the specific tasks necessary to achieve the stated research objectives are not obvious and the level of information required to engage with research questions not absolute. These applied, problem-driven research projects, Sedlmair et al.

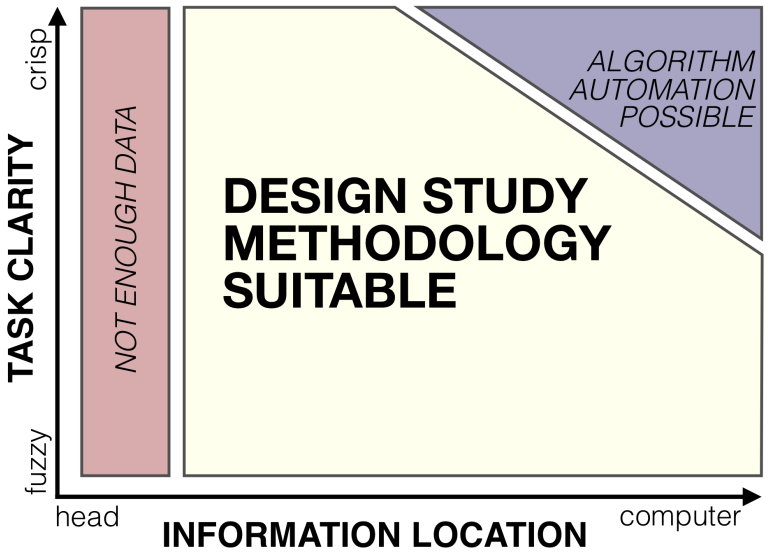

Figure 2.1: Research task and information space as appears in Sedlmair et al. (2012, p. 2433).

(2012) term them *design studies*, are generally exploratory in nature. They start with very specific and real-world problems, but with ambiguity around the research tasks and data models required to engage with these problems, their goal is to progress towards a point where the research tasks and the information used are more obvious and the analysis techniques increasingly automated. This is perhaps best captured in Figure 2.1, taken directly from Sedlmair et al.'s (2012) paper. Two conceptual axes are presented: task clarity and information location. Task clarity depicts how specific, stable and large the tasks necessary to answer a research problem are. Whilst task clarity can be both very precise or more nebulous, Sedlmair et al. (2012) suggest that domain problems are often not clearly defined. Information location characterises the extent to which data and contextual information required to carry out a set of analysis tasks are available to a researcher. In design studies, information location is never perfect. If it were, and if the tasks were well defined, visual analysis methods would not be needed. Instead, an existing set of algorithms might be automatically applied.

In this project, the three research questions are quite broad. Although the raw data might appear concrete (a set of customer and journey records), they alone are not particularly attribute rich and the extent to which modelled and external contextual data might also be used to answer these research questions is more ambiguous. As usage data are explored visually and patterns are discovered within the data, new hypotheses about behaviour are proposed, along with a more specific set of research questions and information tasks.

This movement from the more ambiguous research and information task to the more specific – a necessary contribution of a design study (Sedlmair et al. 2012) – is reflected through the analysis chapters.

The ambitions and research contributions of this research nevertheless do diverge in an important way from Sedlmair et al.'s (2012) conception of a design study:

> A design study is a project in which visualization researchers analyse a specific real-world problem faced by domain experts, design a visualization system that supports this problem, validate the design, and reflect about lessons learned *in order to refine visualization design guidelines*.
>
> (Sedlmair et al. 2012, p.2432; emphasis added)

Under this definition, 'visualization guidelines' perhaps become the object of analysis. By contrast, in this data-driven research it is the domain-specific insights and exposition of the LCHS dataset as a means to studying cycle behaviour, that is the main contribution. This privileging of application domain over computational novelty is justified with recourse to current critiques of the 'computational social sciences': that too often such data-driven studies are overly concerned with computational novelty and scalability across datasets and domains and, as a result, tend to lack analytical substance (Giles 2012, Watts 2013). This research does argue, with evidence, that a visual analysis approach is highly effective where new datasets are used for new purposes. It also demonstrates how visual techniques facilitate exploratory data analysis (Chapters 4 and 6) and support appropriate algorithm selection and development (Chapter 5). The main motivation, however, is not to abstract, refine and therefore contribute a unique set of visual design guidelines. Rather, existing guidelines are applied to a new application area and a case is made for their use in this and other application areas.

2.3 Problematising *design studies*

A number of factors apparently threaten the success of a design study. Many of these arguably apply to all research projects. Starting analysis before defining the research problem, before consulting with domain experts, before learning enough about a domain area and before sufficient data have been provided, are all early problems. So too is a design study that is not informed by literature. Implementing visual analysis techniques

without recourse to visual perception theory and design principles is a problem particular to visualization research (Munzner 2008). Since in many cases a design study may be concerned with exploratory data analysis (Tukey 1977), the process of building analysis tools should be rapid. If *tool building* occupies a significant amount of researcher time, this serves as a brake on analysis, again a pitfall common in visualization research (Sedlmair et al. 2012). Also on the application of visual analysis techniques, Sedlmair et al. (2012) caution against studies that prioritise visualization novelty over domain-specific insights. Finally, Sedlmair et al. (2012) acknowledge that, when conducting a design study project there is rarely a distinct analysis or 'findings' phase. As is often the case in social science research, data insights are usually considered and eventually articulated at the writing phase (Sedlmair et al. 2012).

2.4 Visual analysis in this research

Sedlmair et al.'s (2012) warnings against design studies that are insufficiently grounded within a target domain and overly preoccupied with novel visualization technique, again seem particularly prescient given those same critiques have been levelled at the 'computational social sciences' (Miller 2010, Giles 2012, Watts 2013). In this research, a significant amount of time was spent early on with policy makers and database owners at TfL. The time was used to establish precisely what information could be made available, as well as any possible data protection issues associated with sharing information. A 'research problem' was identified through engaging with the Transport Studies discipline: attending and contributing at conferences, assimilating existing literature on cycling behaviour and also through discussions with policy specialists at TfL. Each analysis chapter starts with a précis of this relevant literature and how the analysis undertaken might contribute to that literature. To ensure that research themes and findings are policy-relevant, meetings with relevant contacts at TfL were also held in order to identify and characterise research needs. In terms of design, visual analysis techniques were implemented only after considering current research in information visualization and visual analytics (Chapter 3); and the programming environment used to develop visual analysis applications (*Processing*) has a set of libraries attached to it specifically designed for rapid prototyping. Finally, by discussing research findings with a wider group of policy makers at TfL, regularly presenting analysis techniques to academic and industry audiences and publishing discrete sets of analysis in relevant academic journals, research findings and their implications were routinely considered rather than simply deferring this sense-making activity to the

final stages of the research project.

2.5 Moving forward

In this chapter, a visual approach to data analysis was briefly outlined. It was argued that visual approaches are particularly suited to this study – both to the dataset being used and to the research ambitions set out in Chapter 1. The application of techniques and approach to analysis in this context is a secondary contribution of this study. Despite the importance of approach, the work described here primarily aims to contribute to the Transport Studies domain. The link between analysis activities and domain-specific contributions is made clear in the analysis chapters. Each starts with a discussion of the research context and literature under which analysis activities are defined and to which contributions are made. Rather than including a high-level literature review chapter, then, relevant studies are incorporated throughout. The findings in Chapter 4, for example, support much of the existing literature on gender and urban cycle behaviour; in Chapter 6 new findings are offered on the subject of group cycling; and in Chapter 5 a novel analysis technique, given other recent analyses of large-scale OD transport datasets, is described. Rather than repeating lengthy descriptions of datasets and techniques in each of these analysis chapters, some time is spent in the following chapter discussing the datasets and derived variables used throughout the analysis, the main visual analysis application and some important underlying design principles.

Chapter 3

Analysis design

Abstract

The data analysis in this study is underpinned by two usage datasets from the London Cycle Hire Scheme (LCHS): a customer database and a full set of journey records from the scheme's inception through to 27th April 2013. Individual cyclists' usage can be interrogated with a unique identifier variable that appears in both these datasets. Around 146,000 bikeshare customers are linked to almost 14 million journeys. The customer database is augmented by incorporating geodemographic classifiers and precomputing behavioural variables for summarising how heavily individuals use the scheme, but also the types of user that they are. Initially, simple cross-tabulations were constructed to identify which geodemographic variables are associated with membership of which behavioural groupings. However, nothing was known about the *nature* of journeys being made by these different groupings: where in the city journeys are made and when they are made. An exploratory visual analysis application was therefore built for performing these sorts of queries 'on the fly'. The visual encodings and comparisons implemented in this application – the use of colour, layout and symbolisation – are informed by current research in flow visualization and established information design principles.

3.1 Datasets

3.1.1 London Cycle Hire Scheme and usage data

The London Cycle Hire Scheme (LCHS) was launched in central London in late July 2010, initially with 5,000 bikes and 315 self-service docking stations. The scheme expanded into east London in March 2012, with the number of bikes increasing to 8,000 and docking stations to 570. In December 2013, there was a further phase of expansion, with a substantial number of new docking stations built in west and south west London. As with most so-called 'third-generation' schemes (Fishman et al. 2013), the LCHS is designed for making short, high frequency and one-way cycle journeys. The first 30 minutes of travel are free; facilities for locking bikes separately from bikeshare docking stations are not provided; and the network of docking stations is dense, with neighbouring docking stations located no more than 300m apart from one another. It is possible to use the scheme either as a formal 'member', or to pay on the day of travel as a 'casual' payment user. Members sign up to the scheme in advance and are issued with keys for quick access to bikes. Casual users insert a payment card at unmanned kiosks located at bikeshare docking stations in order to release a bike.

A complete set of usage data from the scheme's inception through to midnight on 27th April 2013 was provided by TfL. This amounts to 19,822,198 journeys: 5,926,835 made by casual users and 13,895,363 made by members. A condition of the data sharing agreement, established with TfL when planning this research, was that data are stored locally and on an encrypted disk partition. Data are therefore processed and queried using the serverless database engine *SQLite*[1]. Schemas for the three database tables that constitute the LCHS usage data appear on the following page (Tables 3.1, 3.2 and 3.3).

The 'Journeys' table is the largest and stores information on individual journey events: the docking station a bike is undocked from (its origin) and subsequently docked at (its destination), associated timestamps for these events and an identifier representing the individual customer making that journey. The 'Members' table contains information on all individuals who signed up to the scheme as formal members. Here, members' gender (derived from full name and title), 'home' postcode (linked to OSGB Eastings

[1] `http://www.sqlite.org`

and Northings using Code-Point®[1]) and date of registration are recorded. Database managers at TfL themselves coded the gender variable, as information on the full name and title of members cannot be shared. In some instances, 0.6% of the dataset, it was not possible to make these assumptions: for example, an individual with a gender-neutral title (Dr.) and first name (Hilary). Finally, information on the name and geographic location, again recorded as OSGB Eastings and Northings, of docking stations is stored in a 'Stations' table. The analysis discussed here is fundamentally based on queries to, and manipulation of, these three tables – of customers, their journeys and of bikeshare docking stations.

Notice that *individual* cyclists' scheme usage can be identified by linking the Members and Journeys tables on the memberID variable and that only those registered as formal members are stored in the Members table. Whilst in the Journeys table, an entry under memberID will always appear, it is clearly not possible to derive the gender and home postcodes of casual payment users as these individuals do not appear in the Members database. An early decision was therefore to only study the usage behaviours of formally registered members of the scheme. This decision perhaps has the greatest implications for the group cycling analysis described in Chapter 6.

memberID	oTime	dTime	oStation	dStation
####	2010-07-30T09:55	2010-07-30T10:01	308	208
####	2010-07-30T09:56	2013-07-30T10:12	290	286
####	2010-07-30T09:55	2013-07-30T10:05	81	174
####	2010-07-30T09:55	2013-07-30T10:23	14	169
⋮	⋮	⋮	⋮	⋮

Table 3.1: Journeys table schema.

userID	gender	postcode	registrationDate
####	F	EC1V ###	2010-09-05
####	M	SE22 ###	2010-09-05
####	M	N1 ###	2010-09-05
⋮	⋮	⋮	⋮

Table 3.2: Members table schema.

[1] `http://www.ordnancesurvey.co.uk/business-and-government/products/code-point.html`

stationID	name	easting	northing
1	River Street , Clerkenwell	531202	182838
2	Phillimore Gardens, Kensington	525207	179398
3	Christopher Street, Liverpool Street	532984	182007
4	St. Chad's Street, King's Cross	530436	182918
⋮	⋮	⋮	⋮

Table 3.3: Stations table schema.

3.1.2 Data cleaning

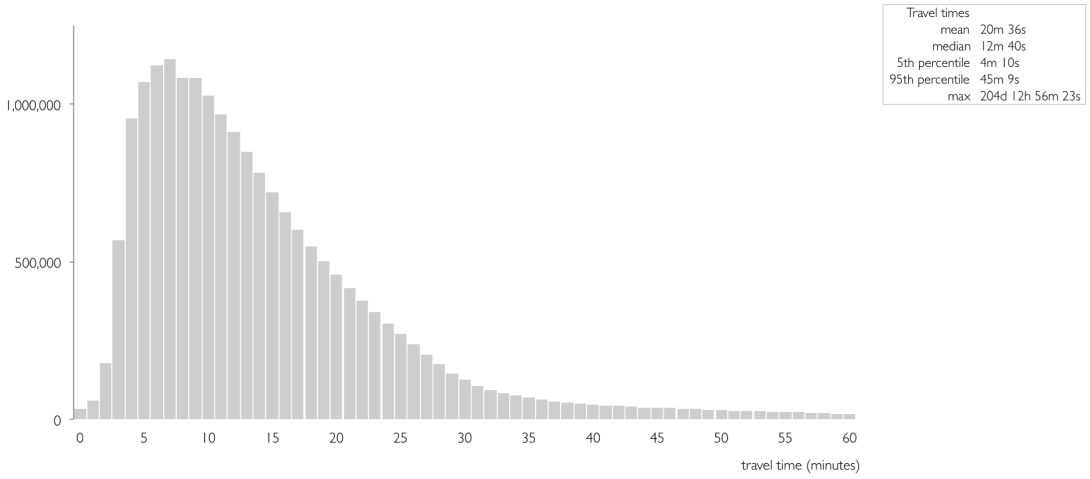

Figure 3.1: Frequency distribution of journeys ordered by observed travel time in minutes. All journeys from the scheme's inception through to 27th April 2013 are presented.

The information in the Journeys table is relatively unproblematic and requires little 'data cleaning'. Studying a frequency distribution of travel times for all cycled journeys, very few, just 10% of these, extend over the 30 minute threshold of free travel and fewer still (4%) are longer than an hour. However, a small number of journeys take under one minute to complete. Since LCHS docking stations are located very close to one another, it is conceivable that users may release a bike from a docking station and cycle, perhaps downhill, to a nearby docking station within 60 seconds. The motivation here might be to shorten the travel time for a journey that otherwise would be taken by foot. It is also likely, though, that many of these short trips finish at the same docking stations they started at and may represent failed hires: a cyclist releases a bike from a docking station or decides against making a bikeshare journey for whatever reason. Whilst such events

may be interesting for service-level analysis, this research is concerned with studying genuine cycle journeys. All instances of 'failed' hires – hires where a bike is released from a docking station and subsequently docked at that station within a one minute window – were removed. This amounts in total to 31,822 journeys, 0.2% of the total number of journeys in the database. As the maximum travel time and difference between mean and median travel times suggests (see Figure 3.1), there are also a small number of unreasonably long hire durations in the journeys database. Since these are very rare, they are unlikely to impact on research findings. However, to ensure that only genuine bikeshare hires are analysed, journeys that exceed the maximum hire duration permitted through the scheme before a £150 penalty is levied (24 hours) were also cleaned out. In total this applies to 9,245 journeys made since the scheme's inception through to 27th April 2013.

3.1.3 Geodemographics

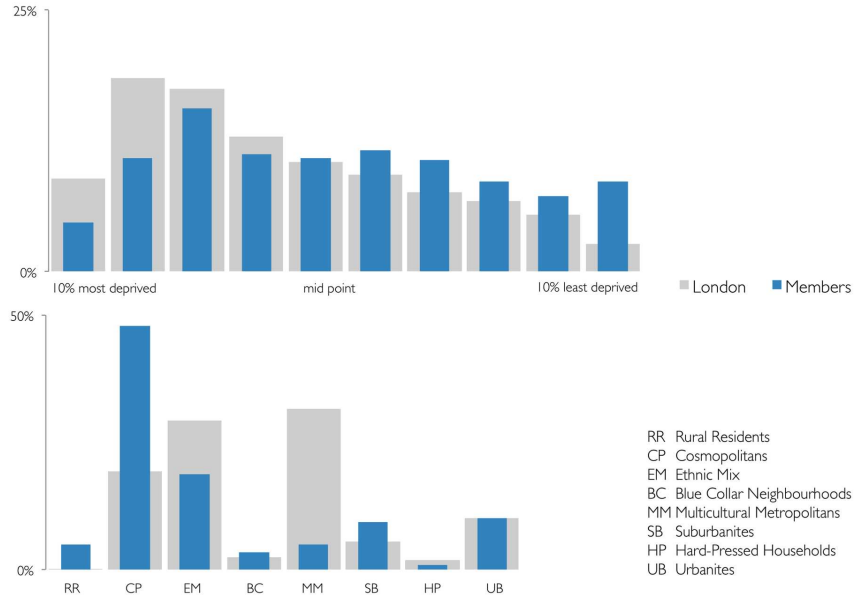

Figure 3.2: Percentage of members in each IMD decile (above) and OAC group (below) compared to that of London. There are greater relative numbers in the least deprived IMD deciles and the Rural Residents, Cosmopolitan and Suburbanites OAC groups than might be expected given the population profile of the London Region (based on 2011 Census). Members making journeys between 14th September 2011 - 2012 are presented.

This work aims to identify, describe, suggest likely motivations for and, if possible, explain the cycling behaviours of bikeshare customers. Existing research suggests that attitudes towards cycling often have a distinct gendered (Gatersleben & Appleton 2007, Garrard et al. 2008) and social-demographic (Anable et al. 2010, Buehler & Pucher 2012, Goodman 2013) expression. The only demographic information directly available on members is their gender. Also available, though, is the postcode that members registered with. This is the address to which keys for accessing bikes are mailed once registering and it must match the full name provided on members' payment cards. Most likely this is a home address, though clearly in some instances customers may have registered with a work payment card and address, or their card may be linked to an address at which they no longer live permanently. Assuming that it does represent a customers' current home, the postcode variable was linked to two freely available geodemographic datasets: the preliminary Census 2011 Output Area Classification (OAC)[1] and the 2010 Indices of Multiple Deprivation (IMD) (Department for Communities and Local Government 2011). Unit postcodes were matched to Output Areas using the Office for National Statistics Postcode Directory[2]. For every member with a valid postcode, fields recording an OAC group and IMD score, and therefore some indication of the types of communities in which customers apparently live, were added to the customer database. In Figure 3.2, the OAC and IMD profiles of members are compared to that of London's population. It should be stated here that OAC and IMD are area-level measures and there is no guarantee that individual LCHS cyclists share the same characteristics as their immediate neighbours.

[1] http://www.opendataprofiler.com/2011OAC.aspx

[2] http://www.ons.gov.uk/ons/guide-method/geography/products/postcode-directories/-nspp-/index.html

3.1.4 Distance from home to docking station

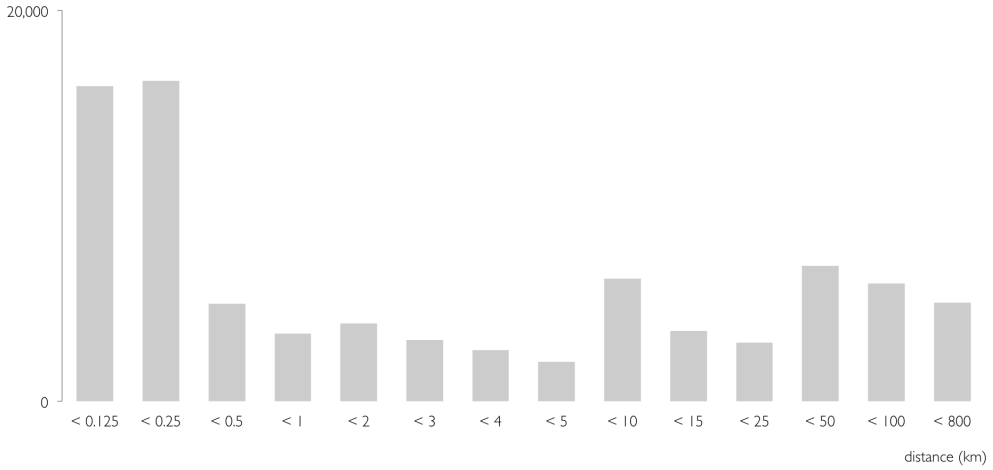

Figure 3.3: Straight-line distance to nearest docking station for members. Assuming postcodes do represent members' home addresses, 45% apparently live less than 0.5km from a docking station, but 22% also live more than 25km from their nearest station. Members making journeys between 14th September 2011 - 2012 are presented.

In their household survey of users and potential users of Montréal's bikeshare scheme, Fuller et al. (2011) found much higher rates of usage for residents living within 250m of a docking station than those living further distances from the scheme. How far members need to travel from home to reach a LCHS bike might also affect *how* those individuals use the scheme and the types journeys they make. Since geographic coordinates are stored for members' 'home' postcodes, it is possible to calculate straight-line distances from customers' homes to their nearest docking station. Variables identifying this distance from home to nearest docking station, and recording the stationID this represents, were therefore added to the Members table for every customer with a correctly geocoded postcode (just 0.02% of customers' postcodes could not be geocoded).

3.2 Behavioural variables

3.2.1 Personalised travel times

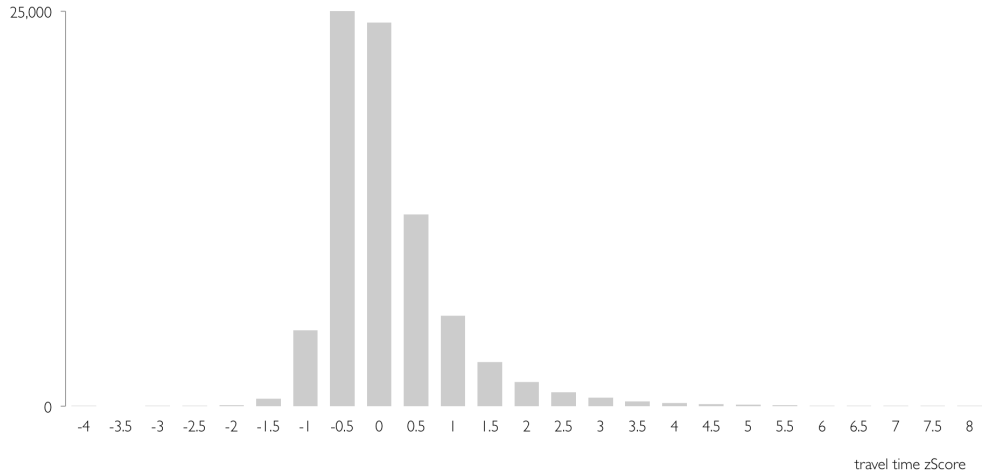

Figure 3.4: Travel time $z-scores$ for members. Members making journeys between 14th September 2011 - 2012 are presented.

The motivation behind accessing the customer database was to link individual customers with their journeys and describe and explore how customers variously use the scheme. One factor, which might discriminate different types of user or usage characteristic, is relative trip duration. For every customer with a valid set of journeys, personalised travel time $z-scores$ were calculated, where the travel time for every customer journey (u_{od}) is compared to the average travel time for that OD pair made by the total population of bikeshare customers ($\overline{p_{od}}$). Following statistical probability theory (Bartholomew et al. 2008), this score was only computed where the total frequency for that journey pair made by the member population was at least 30 ($f_{od} >= 30$):

$$z - score(u_{od}) = \frac{u_{od} - \overline{p_{od}}_{(fod>=30)}}{std(p_{od(fod>=30)})}$$

Users' travel times were, then, only benchmarked against population travel times where there was reasonable confidence in the population mean and clearly only against corresponding journeys (OD combinations). A minimum journey time of 3 minutes and a maximum of 4 hours was also used when computing the scores. The average of each users' set of $z-scores$ was taken in order to give a single score for each member and since travel

time distributions are generally positively skewed, the scores were made to fit a normal distribution by taking the square root of travel times when performing the $z - score$ calculation. In Figure 3.4, members with faster travel times for the journeys they make have negative $z - scores$; those with slower travel times have positive $z - scores$.

3.2.2 Recency-Frequency segmentation

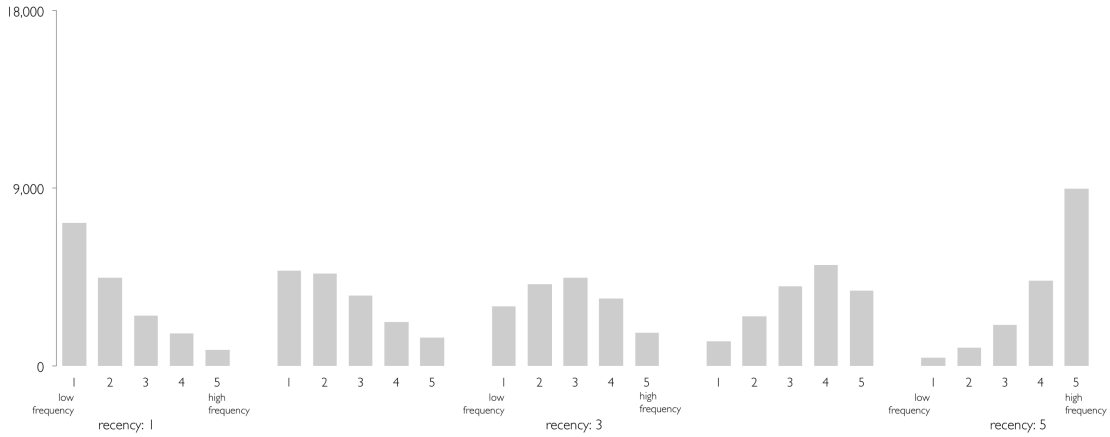

Figure 3.5: RF segmentation for members. Whilst 11% of members can be found in the top RF segment (heavy and active scheme users), 9% are in the bottom RF segment (typically using it once or twice after subscribing to the scheme). RF segmentation for members making journeys between 14th September 2011 - 2012 are presented.

Recency-Frequency (RF) segmentation is a very simple technique used in direct marketing to classify and group together similar customer purchase behaviours (Kohavi & Parekh 2004, Novo 2004). It is based on empirical research that finds Recency - how recently a customer bought or used a product - to be a strong predictor of how likely that individual is to buy a product again soon (Kohavi & Parekh 2004). Frequency - how often a customer buys or uses a product - is also a predictor of future purchase, but combining both scores together helps further discriminate purchase behaviours. Whilst more sophisticated techniques exist, RF segmentation provides a simple technique that can be applied to most action-oriented customer datasets (Novo 2004).

The Recency measure was calculated by identifying customers' most recent journey, ordering customers according to this most recent journey and assigning discrete scores within five equal frequency bins, from most (score 5) to least (score 1) recent. For Frequency, the first and last journey appearing in customers' records were identified and the

total number of journeys that customer made was divided by the total time (in days) that elapsed between these dates. After exploring these derived Frequency scores against the actual number of journeys made by each member, members completing several journeys within a single 24 hour period appeared to have an unduly high Frequency score. A minimum duration of 90 days was therefore imposed when calculating the Frequency measure. Combining the two scores gives 25 customer segments and in Figure 3.5 these are represented as a set of column charts.

It is common in RF analysis to have significant numbers at both extremes of Recency and Frequency scores (as in Figure 3.5) and as a relatively crude segmentation technique, each classification group is not entirely homogenous. As Recency scores increase, the gap between the middle point of each recency score decreases and as Frequency scores increase, there is some variation in absolute frequency values (see Figure 3.6). Members making extremely large numbers of journeys become a problem when querying smaller subsets of the population; users that make very large numbers of the same journey (same OD pair) can dominate and there is a risk of generalising the spatial travel behaviours of the wider bikeshare population based on the travel patterns of these extreme users. This was an observation made whilst exploring usage behaviours in the visual analysis software described in the following section (Section 3.3). Studying a further distribution of journeys made by the top 1,000 most prolific members, the 98 LCHS customers that made more than 730 journeys over a 12-month study period were excluded from the analysis described in this study. The behavioural variables discussed above were also recalculated excluding these 'outlier' members.

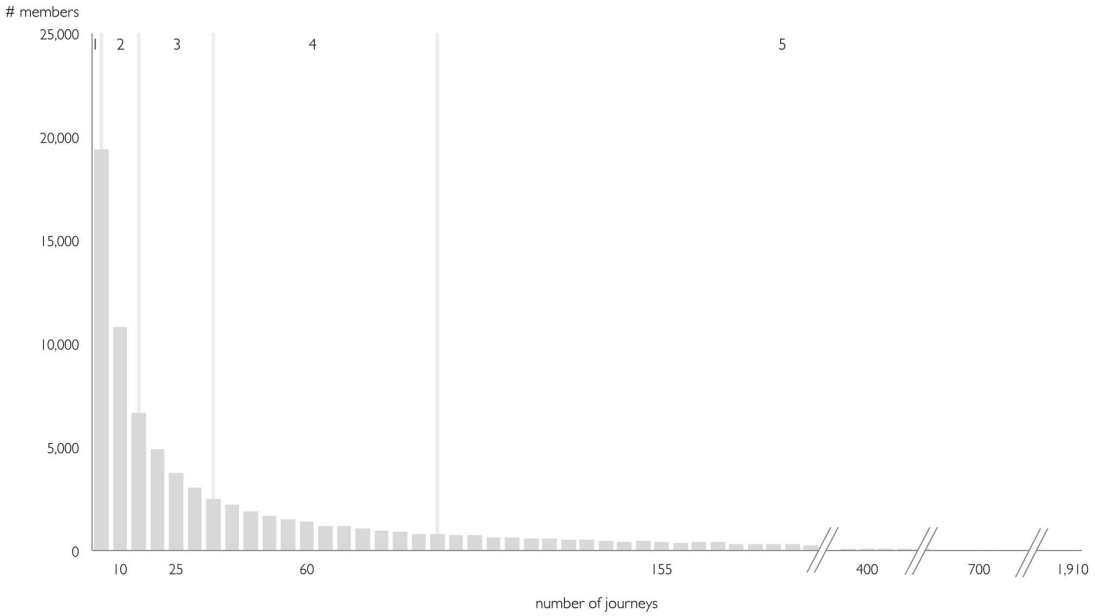

Figure 3.6: Frequency distribution of members ordered according to the absolute number of journeys they have made. Members making journeys between 14th September 2011 - 2012 are presented.

3.2.3 Temporal clustering

Whilst RF segmentation provides a useful means of quickly identifying regular and/or active members from those that are less active, or no longer active members, it does not provide a summary of the *types* of journeys that members typically make. As might be expected for a shared transport system, early analysis of LCHS usage data found very distinct spatial usage behaviours associated with specific times of day and days of the week (Wood et al. 2011, Lathia et al. 2012). A technique was therefore also developed for automatically identifying and summarising groups of members who typically use the LCHS at particular times of day. Such an approach was taken by Lathia et al. (2013) when analysing a sample of OD smartcard data from the London underground network. Here, each traveller was represented as a vector of values summarising when they travel and agglomerative *hierarchical cluster analysis* (HCA) used to identify groups of customers sharing similar temporal usage profiles.

Following Lathia et al. (2013), five input variables were created for the temporal clustering: morning peaks (weekdays between 6am-9am), evening peaks (weekdays between 4pm-6.30pm), interpeaks (weekdays between 10am-3pm), evenings (weekdays or week-

ends between 9pm-11pm) and weekends (between 8am-9pm). A random sample of 2,000 members was drawn and a Euclidean distance matrix constructed where, for a given pair of members i and j, the distances between each variable k – between the values a customer has in each time bin – was computed:

$$\delta_{i,j} = \sqrt{\sum_{k=1}(x_{ik} - x_{jk})^2}.$$

Once the distance matrix was constructed, the *Ward's* method (Bartholomew et al. 2008) was used for identifying and merging together similar members. It should be noted that there are other methods for performing this agglomeration of objects. In the nearest neighbour or single linkage method, the closest pair of objects in the n^2 distance matrix is found and merged to form a new cluster. A new distance matrix is then calculated before the next two closest objects are identified and merged. In the furthest neighbour or complete linkage method, the approach is very similar, but the difference between two agglomerated groups is defined as the distance between the most distant neighbour in each of these groups (Bartholomew et al. 2008). The *Ward's* method considers all pairs of objects and establishes how much information would be 'lost' – defined as the sum of squares about the mean of a cluster centre – if the pair were to be merged. The pair that is merged always involves the least loss of information (Bartholomew et al. 2008). HCA was trialled using each of these agglomeration methods. Visually inspecting the output dendrogram from this analysis, as well as Average Silhouette Width (ASW) values (Rousseeuw 1987), calculated at different cuts of the dendrogram, a 5-cluster solution using *Ward's* agglomeration resulted in the most stable and coherent clustering (ASW: 0.40).

Since HCA involves constructing a distance matrix where n^2 objects are compared exhaustively for the highest similarity, it is computationally expensive and cannot be extended beyond the 2,000 sample of members. $K - means$ clustering was therefore used to run this analysis on the full member population. Unlike HCA, $k - means$ requires an appropriate number of output clusters (k) to be first specified. Based on the initial HCA, a 5-cluster solution was specified when executing the $k - means$ analysis. In order to improve the stability of the solution, the algorithm was run with 100-random starts and the optimum solution was selected based on the maximum intra-cluster similarity and inter-cluster difference between objects.

The five output clusters are presented in Figure 3.7. As with the RF segmentation, there will be some internal variation within cluster groupings. However, the cluster memberships provide a useful, data-driven means of summarising customers according to the 'types' of user that they are. An important observation, for instance, is that a substantial portion of LCHS customers (27%) are so-called 'anytime users': relatively heavy scheme users who make a diverse set of journeys and apparently use the scheme for both leisure and commuter purposes. In their comparative analysis of China's major bikeshare schemes, Yang et al. (2011) studied the nature of scheme usage by asking cyclists to recall the purpose of their *most recent* journey. This 'journey purpose' variable was then used to make wider claims around how bikeshare facilities fit within individuals' travel options and as a contextual variable for further analysis. With full and accurate historical data on individuals' scheme usage, a more sophisticated profile of individual-level usage is generated using the clustering procedure described above. For example, the important 'anytime users' group would be entirely missed using Yang et al.'s (2011) method and the corollary – applying the same technique proposed here but with Yang et al. (2011)'s survey data – would clearly be problematic; it would require respondents to accurately recall the specific day and time of every bikeshare journey they have made.

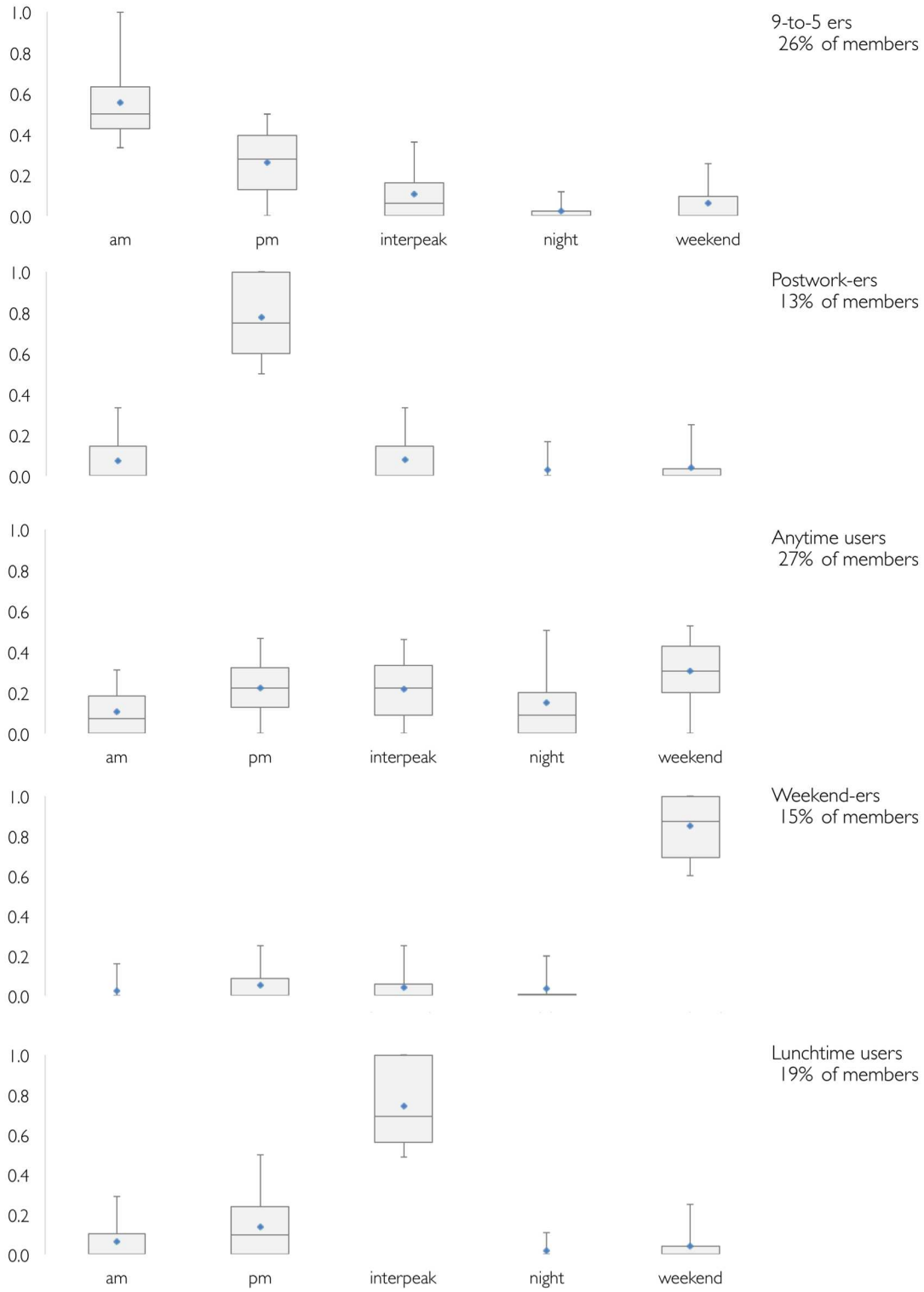

Figure 3.7: Box plots summarising the temporal profile of members in each cluster grouping. For each member, the number of journeys made in each time bin is expressed as a percentage of that member's total journeys. Cluster groupings for members making journeys between 14th September 2011 - 2012 are presented.

3.2.4 Analysis period

Note that the summaries of members appearing in the figures are based on a 12-month period of usage. The motivation for selecting a consistent 12-month period is discussed in Chapter 4 and the September 2011-2012 period was chosen purely because this represented the most recent set of membership data available when the majority of analysis work was conducted. Very recently, access has been given to usage data through to April 2013. An aspect of research, made possible by the fact that LCHS data are recorded continuously, is that of change over time. As the scheme expands and becomes more established, it is likely that new types of customer are introduced to the scheme, but also that the behaviours of existing customers might shift. Detailed analysis of these changes is beyond the scope of this research. However, in Appendix B, the customer-related segmentations and exploratory analysis discussed in this chapter are re-run to identify behaviours that remain consistent and those that appear to have changed.

3.3 Visual analysis

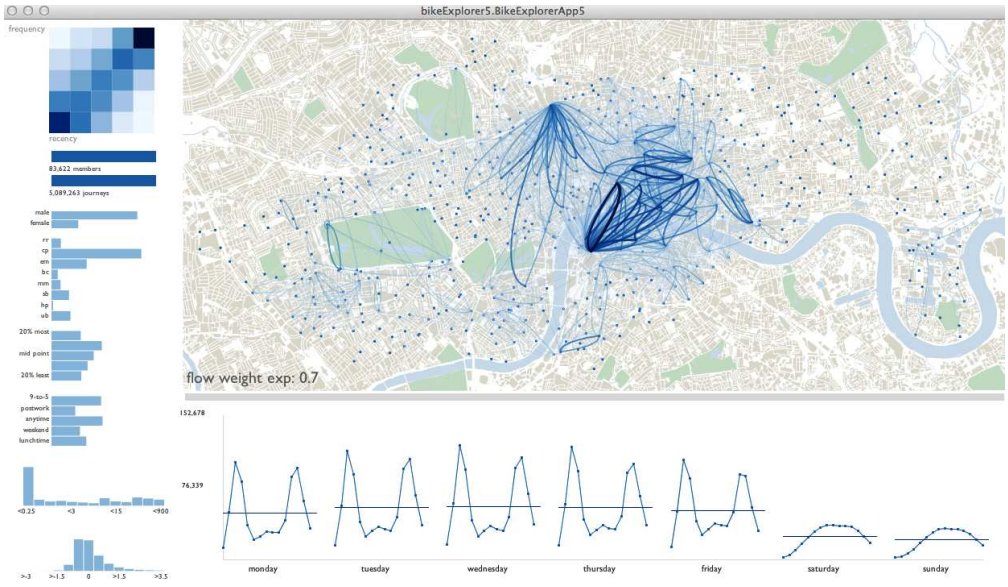

Figure 3.8: Visual analysis application combines a spatial (centre), temporal (bottom) and customer-related (left margin) view. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

An early analysis objective, after computing the behavioural classifications, was to explore the extent to which levels and types of usage, as defined by the RF segmentation and temporal cluster variables, vary by geodemographic and other derived variables. Querying these data, cross-tabulations using the Chi-statistic were computed, whereby observed frequencies within, for example, each RF segment for a subset of members were compared to modelled (expected) frequencies given the member population as a whole. This exploratory analysis enabled differences in the level of scheme usage to be related to customers' gender, geodemographic classification, user type and how far customers apparently live from their nearest docking station. It was, however, increasingly necessary to explore, test and compare *multiple* combinations of these derived variables simultaneously. Also, nothing was known about the temporal or spatial structure of the journeys being analysed. The very early analysis approaches - querying the dataset within *SQLite* and making simple calculations within the statistical analysis environment *R*[1] - frustrated these more detailed analysis requirements. A visual analysis application for performing these sorts of queries 'on the fly' was developed in *Processing*[2], a Java based programming environment often used for developing visual analysis software. This application combines three coordinated and linked views (Dykes 1997, Roberts 2005), enabling a spatial, temporal and customer related summary of members' cycling behaviours (Figure 3.8). Some time is spent here discussing each of these views in turn: the use of visual encoding and symbolisation is critically discussed, along with details on the nature and level of interactions the application enables.

3.3.1 Spatial overview

To show the spatial structure of members' journeys, lines between all possible journey (OD) pairs are drawn. This is done using Bezier curves and following Wood et al. (2011), direction is encoded by making these curves asymmetric – the straight end representing journey origin, the curved end journey destination (Figure 3.9). To overcome problems of visual clutter and salience bias common in flow visualizations, Wood et al. (2011) proposed a weighting factor that emphasises flow magnitudes. This same weighting factor (w_{od}) is used in the application developed in this research where, for each unique OD pair, the number of journeys made between that pair of docking stations (f_{od}) is scaled to the most frequently travelled OD pair in the dataset (f_{max}):

[1] `http://www.r-project.org`
[2] `http://www.processing.org`

$$w_{od} = \left(\frac{f_{od}}{f_{max}} \right)^{exp.}$$

The weighting factor determines the thickness, transparency and colour of each flow line so that there is a direct mapping between flow frequency and visual saliency. Varying the exponent (*exp*) alters the impact of the weighting factor and if decreased, allows less common flows to be given slightly greater prominence. Finally, to further ensure that less common flows do not occlude more common ones, OD pairs are ordered from least to most frequent and then drawn in this order. It is possible to alter the exponent used in this weighting factor and increase or decrease the saliency given to the less frequent flows. Also, by clicking and dragging, or rather brushing (Becker & Cleveland 1987), the flow map spatial selections can be performed (Figure 3.9): journeys made in particular parts of the city can be selected and a view of when (see 3.3.2) and by whom (see 3.3.3) those journeys are made appears.

Cartographers have for some time struggled with techniques for best representing flow data and alternative means for representing journeys certainly exist. Working with a month's journey data from the LCHS, Wood et al. (2011) discussed these problems in detail. The principal challenge is in dealing with flows that are both large in scale and spatially complex. Directly mapping large numbers of journeys by drawing lines between OD pairs leads to a cluttered graphical display, with any structure almost unintelligible (Wood et al. 2011). Data aggregation or reduction may overcome these problems; so too might alternative visual representations or more subtle manipulation of visual variables (Bertin 2010) used to represent flows.

The techniques implemented above, reading Wood et al. (2011), involve manipulating the visual variables used to encode flows and overcome the main problem - of visual clutter. However, the problem of salience bias, where longer flows obscure shorter but possibly important flows, is only partially overcome by drawing flows from least to most frequent. Alternative depictions aimed at further reducing salience bias and enabling a more detailed analysis of journeys, specifically Wood, Slingsby & Dykes's (2010) spatially ordered OD map, were considered. However, the visual metaphors used in the flow line symbolisation can perhaps be more quickly interpreted than the spatially ordered OD matrix and are therefore better suited to exploratory querying.

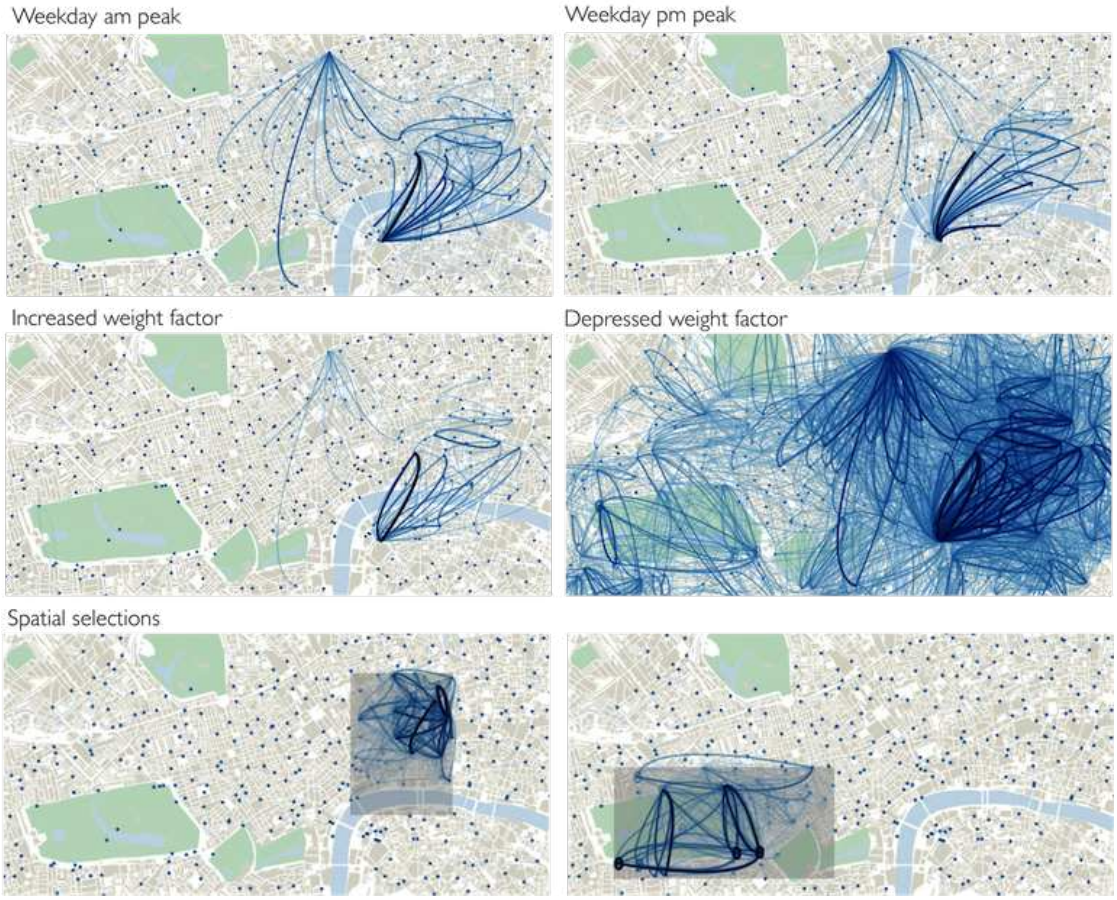

Figure 3.9: Direction of journeys is shown by making the origins of flow lines straight and destinations curved (top). The emphasis given to journey frequency can be varied and journeys in particular parts of the city can be selected with mouse interaction. Colour values are chosen from the Brewer 'Blues' sequential colour scheme (Harrower & Brewer 2003). Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

3.3.2 Temporal overview

The temporal view displays hourly daytime usage by day of week. It is possible to query journeys that are made at particular times of day, days of week and, using a temporal slider, analyse behaviours over varying temporal resolutions. The decision to aggregate journey volumes at hourly intervals was made whilst reviewing approaches others have taken when interrogating similar datasets (Blythe & Bryan 2007). Since the aim of the temporal view is an immediate structural overview, hourly aggregation intuitively makes sense. The graphic itself - a variant of the cycle plot (Robbins 2005) - enables analysis of hourly flows, but also rapid comparison of day-of-the week trends; the horizontal line

running through the chart represents the average hourly flows for that day. By overlaying selected subsets of members (blue) with the total member population (grey), it is possible to quickly make comparisons and identify deviations from an expected temporal pattern of scheme usage. This technique is spatially efficient, requires little cognitive effort and is appropriate where comparisons between a limited number of metrics are made (Gleicher et al. 2011).

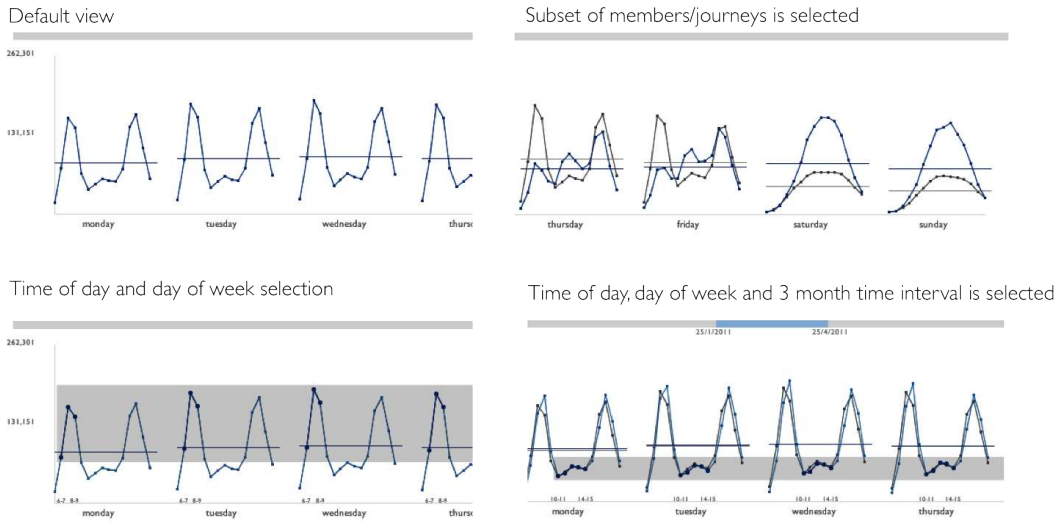

Figure 3.10: Temporal view and possible interactions.

3.3.3 Customer related view

The left margin of the graphic displays the customer related variables. Recency-Frequency scores are presented within a matrix (Kohavi & Parekh 2004, Wood, Radburn & Dykes 2010), the gender, geodemographic and cluster variables appear as horizontal bars and the 'distance to docking station' and travel time $z - score$ variables are shown as histograms.

The matrix view is an efficient way of representing the 25 RF segments. As Figure 3.5 demonstrates, an association between Recency and Frequency is to be expected: members using the LCHS often are likely to have used it recently. Ordering the columns and rows by Recency and Frequency means that similar customer classifications are located near to one another (Friendly 2009). Since position within the matrix is already used to delineate RF category, the number of customers in each RF segment is encoded using

colour lightness and where a subset of users and/or journeys is selected, variation from an expected number of members within each RF segment is shown by varying colour hue (Figure 3.11).

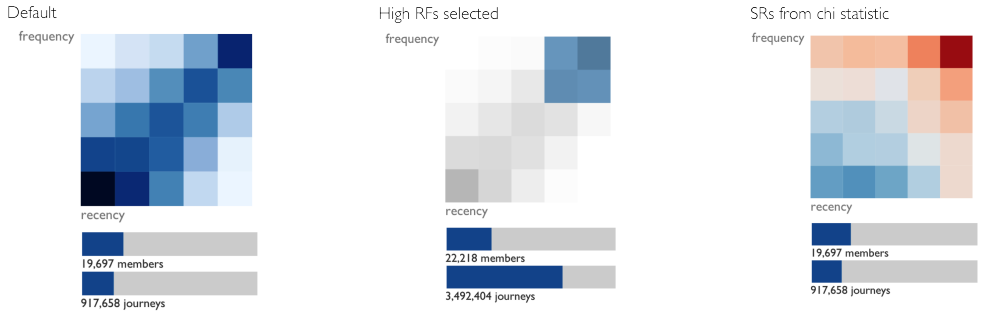

Figure 3.11: RF view and interactions. In the third view, observed numbers of customers in each RF segment are compared to an expected model based on the member population as a whole. Signed Pearson's residuals from the Chi-statistic are mapped onto a Brewer diverging colour scheme (Harrower & Brewer 2003).

The choice of horizontal bars for the gender, geodemographic and cluster variables and vertical bars for the 'distance to docking station' and travel time $z-score$ variables – variables that might be expected to follow a distribution – is logical and efficient. The visual encoding of comparisons is made consistent with the temporal view: selected subsets (blue) are compared to the total member population (blue) by overlaying one on top of the other (Figure 3.12).

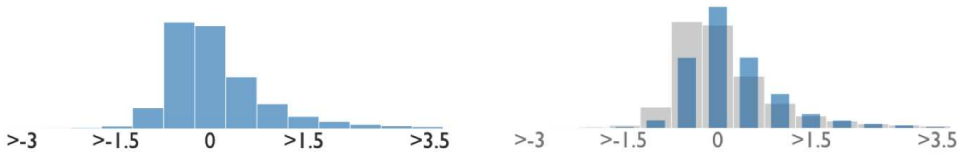

Figure 3.12: Histograms displaying members' travel time $z-scores$. When a selection is performed (right) the relative number in each $z-score$ bin for the selected subset appears in blue; the overall proportion in grey.

3.3.4 Interactions

It is possible to make any combination of spatial, temporal and customer-related selections simultaneously. Simply clicking or dragging on particular geodemographic or behavioural groups, time periods and spatial areas filters those members and their journeys. Additional summary statistics also appear when specific keys are pressed. Ma-

nipulating data in this way means that additional buttons and drop-down menus, which might reduce the 'data:ink ratio' (Tufte 1986), are not needed. There are nevertheless limitations to the interactions. Smooth, animated transitions (Heer & Robertson 2007) between views would better enable changes in the colour, size and position of chart elements to be detected and when making comparisons, a dynamic benchmark would offer greater analytical potential than comparing only against the total member population. These improvements would require substantially more programming time and conscious of Sedlmair et al.'s (2012) prescriptions that visual analysis tool building should be just one aspect of a successful *design study*, time was instead focussed on the insights that could be derived from using this application for exploratory analysis.

3.4 Moving forward

In this chapter, the LCHS datasets were covered in some detail: how data were initially processed and structured, external information used to augment the customer dataset and new behavioural variables that were derived through mining the Journeys data. Decisions for selecting these derived variables were informed by research around broader cycling behaviour (Anable et al. 2010, Buehler & Pucher 2012), of bikeshare usage in other cities (Fuller et al. 2011), as well as approaches taken elsewhere with similarly structured OD data (Lathia et al. 2013). The main visual analysis application was also described and, following Sedlmair et al. (2012), design decisions justified with recourse to accepted design principles and visual perception theory. The derived variables provide important contextual data for exploratory analysis. Combined with the visual analysis software, they enable relatively detailed behaviours to be queried and described. The analysis that immediately follows is based entirely on queries made within the visual analysis application described in Section 3.3 and using these variables. The chapter demonstrates how the software and derived variables enable rich, data-driven hypotheses to be suggested and, through interaction with the software, behavioural and demographic controls to be very quickly explored. As with each of the analysis chapters, Chapter 4 starts with a discussion of relevant domain literature.

Chapter 4

Exploring gendered cycle behaviours

Abstract
In this chapter, men's and women's usage behaviour are explored using the main visual analysis software. Female customers' usage characteristics are found to be demonstrably different from those of male customers. Usage at weekends and within London's parks characterises women's journeys, whereas for men, a commuting function is more clearly identified. Some of this variation is explained by geodemographic differences and by an atypical period of usage during the first 3 months after the scheme's launch. Controlling for each of these variables, by performing various spatial, temporal and customer-related selections, brings some convergence between men and women. However, many differences are preserved. There is a sense that, even when making apparently utilitarian cycle trips, journeys within parts of the city that contain multi-lane roads are comparatively rare and instead female cyclists preferentially select areas associated with slower traffic streets, with cycle lanes slightly offset from major roads.

Perhaps the most interesting and substantial differences identified in this chapter relate to spatial and temporal aspects of men's and women's usage behaviours. Observations around this different spatiotemporal structure are compelling because the LCHS dataset is very large and complete: it contains information on a full population of over 80,000 users cycling within a relatively compact urban area. Such analysis of detailed travel behaviours at the scale of the city would not be possible with the much smaller samples achieved in more traditional, actively collected

datasets.

This work has been published in: Beecham, R. & Wood, J. (2014) Exploring gendered cycling behaviours within a large-scale behavioural dataset. *Transportation Planning and Technology*, 37(1), pp.83-97.

4.1 Research context

The Introduction chapter alluded to the fact that research into the motivations and barriers surrounding cycling is burgeoning (Pucher & Buehler 2012). A substantial aspect of this research relates to gender and cycling behaviour. The reasons for this gendered focus are best enumerated by Garrard et al. (2012). In bicycle-friendly cities and countries, cycling is apparently seen as a highly inclusive activity open to most demographic groups, with rates of female cycling matching or even surpassing that of men (Garrard et al. 2012). In car-oriented cities with low levels of cycling, however, cycling is perhaps seen as the preserve of largely young or middle-aged men (Garrard et al. 2012). Garrard et al. (2012) add that this link between gender and urban cycling is so marked that some (Barker 2009) have suggested the relative balance of men and women cycling in a city might be a proxy for how cycle-friendly that city is.

There are various explanations for the observed gender gap in cycling uptake. Detailed qualitative studies have linked motivations around cycling amongst women and men to particular personal circumstances and life stages (Bonham & Wilson 2012). Larger survey-based research has suggested that differences between men's and women's uptake might relate to preference: men are more likely than women to agree that they enjoy cycling (Emond et al. 2009). A substantial barrier is that of perceived personal safety. A relatively large survey of 1,862 cyclists in Queensland, Australia found that women are more likely to cycle off-road than men, are less likely to commute by bicycle than men and that, although factors related to traffic conditions, motorist aggression and safety are concerns for both women and men, women report a far greater number of these as constraints (Heesch et al. 2012). Similar findings were identified by Tilahun et al. (2007) in a study of participants' stated preferences around route choice. Observational studies have also shown these preferences to be expressed in women's 'real' cycle behaviours. In Portland, Oregon, a sample of 166 self-selected participants were recruited and their

cycling monitored using GPS (Dill & Gliebe 2008). Compared with male participants, women made a smaller share of their journeys on major roads or routes without bike lanes and more often cycled on low-traffic streets or boulevards (Dill & Gliebe 2008).

Given the importance of gender in researching urban cycling, then, this first analysis chapter focusses on men's and women's use of the LCHS. Using the visual analysis software introduced in Chapter 3, special attention is paid to how *spatiotemporal* cycling behaviours differ between male and female bikeshare cyclists. The chapter concludes by further reflecting on findings and the LCHS dataset in the context of existing research.

4.2 Presenting results

In this chapter, visual patterns that emerge from exploring the LCHS dataset are described, but also where possible quantitative evidence is provided to confirm or question these findings. When comparing, for instance, frequencies of men and women in each RF group, contingency tables are created and Pearson's residuals from the Chi-statistic calculated to compare observed frequencies against what would be expected given equality of proportions between men and women:

$$\chi = \frac{obs - exp}{\sqrt{exp}}$$

Pearson's residuals are effectively $z - statistics$ – they are critical values that can be related to $p - values$ – and, combined with an overall Chi-statistic (χ^2), can be used to assess category-level statistical significance (Field et al. 2012). Where only one category is considered – for example, the relative number of men and women making journeys during peak times – a two-way contingency table is created, with an overall Chi-statistic (χ^2) and again signed Pearson's residuals to evaluate the direction of differences.

A problem with using formal significance testing on the bikeshare dataset is that statistical significance is a function both of the real difference between values – the size of the effect – and the size of the dataset from which those values are drawn:

$$\text{statistical significance} = \text{effect size}^* \text{ x sample size}$$
$$^*\text{given variability}$$

As dataset size increases, so too does the statistical power or sensitivity of the test; the size of the effect required to achieve a statistically significant result reduces. With very large datasets, such as the LCHS data, trivially small differences can be labelled as statistically significant. This is not to say that the statistical significance testing is wrong. For the example above, the null hypothesis (H_0) would be that there is no difference between the relative number of men and women who have made peaktime journeys versus those who have not. The $p - value$ from this test would correctly describe the probability of the observed or larger differences between men and women, given H_0 – given an assumption that the data are drawn from a population in which there is no difference between men's and women's peaktime journeys. However, since it is so compounded with sample size, this probability is not particularly useful in the (fictional) example above. It is possible to reject the null hypothesis with a trivially small difference, or effect.

	observed		%		expected		Pearson's resid.	
gender	peak	non-peak	peak	non-peak	peak	non-peak	peak	non-peak
male	55,000	5,000	91.7	8.3	54,900	5,100	0.4	−1.4
female	18,200	1,800	91	9	18,300	1,700	−0.74	2.4

$\chi^2 = 8.6 (p < 0.001 = 7.8)$

Table 4.1: Fictional example of contingency table comparing number of men and women who have made peaktime journeys.

One approach, already used in the previous chapter, is to represent differences graphically (Tukey 1977). Another, is to calculate the *effect size*: a scale-free measure that indicates the magnitude of difference between phenomena of interest (Coe 2002, Cohen 1994, Cohen 1990). There are various ways of quantifying effect sizes. Cohen's d is used for comparing two sample means (see Chapter 7). Cohen's d is simply the difference between two means divided by their pooled standard deviations. It measures the observed difference given the level of variability in a collection of data:

$$d. = \frac{\bar{x}_1 - \bar{x}_2}{SD_{pooled}}$$

In Chi-square, Cramer's V (ϕ_c) is used. This is calculated by taking the square root of the Chi-statistic (χ^2) divided by the sample size:

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

Once calculated, a decision still has to be made about ϕ_c values. Whilst their importance varies with research context, Cohen (1990) suggested that ϕ_c values of 0.1, 0.3 and 0.5 represent small but non-trivial, medium and large effects respectively; for Cohen's d, values of 0.2, 0.5 and 0.8 correspond with small, medium and large effects.

Confidence intervals are another means of interpreting the difference between values in a statistical test, especially since they relate directly to the initial units of measurement (Cohen 1994). However, they too are partly a function of sample size. A second measure of effect size, used in this chapter for two-way comparisons, is the relative risk ratio (RR). This is the probability of an event occurring (a member who has made a peaktime journey) in one group divided by the probability of that same event occurring in a different group:

$$RR = \frac{P_{men}(peak|all)}{P_{women}(peak|all)}$$

This measure is very straight-forward to interpret. If 50% of men had made peaktime journeys, but this was only the case for 20% of women, RR would be 2.5: men are two and one half times more likely than women to make a peaktime journey.

As researchers working in applied domains have access to increasingly large or 'big' datasets, the deficiencies of null hypothesis significance testing (NHST) are particularly prescient. The approach taken here, of paying greater attention to effect sizes, is increasingly common, with several journals in psychology formally requiring authors to report effect sizes alongside inferential tests. Bayesian techniques have also been suggested as an alternative to NHST (Wagenmakers 2007, Kruschke 2013). Within information visualization, Wickham et al. (2010) have introduced the idea of *graphical inference* as a potential approach to performing inferential tests and avoiding false positives that might be introduced when analysts interpret patterns in data graphics *visually*. An open challenge remains, especially for those working in applied data analysis, as to which of these approaches are most appropriate to particular analysis contexts.

4.3 Analysis

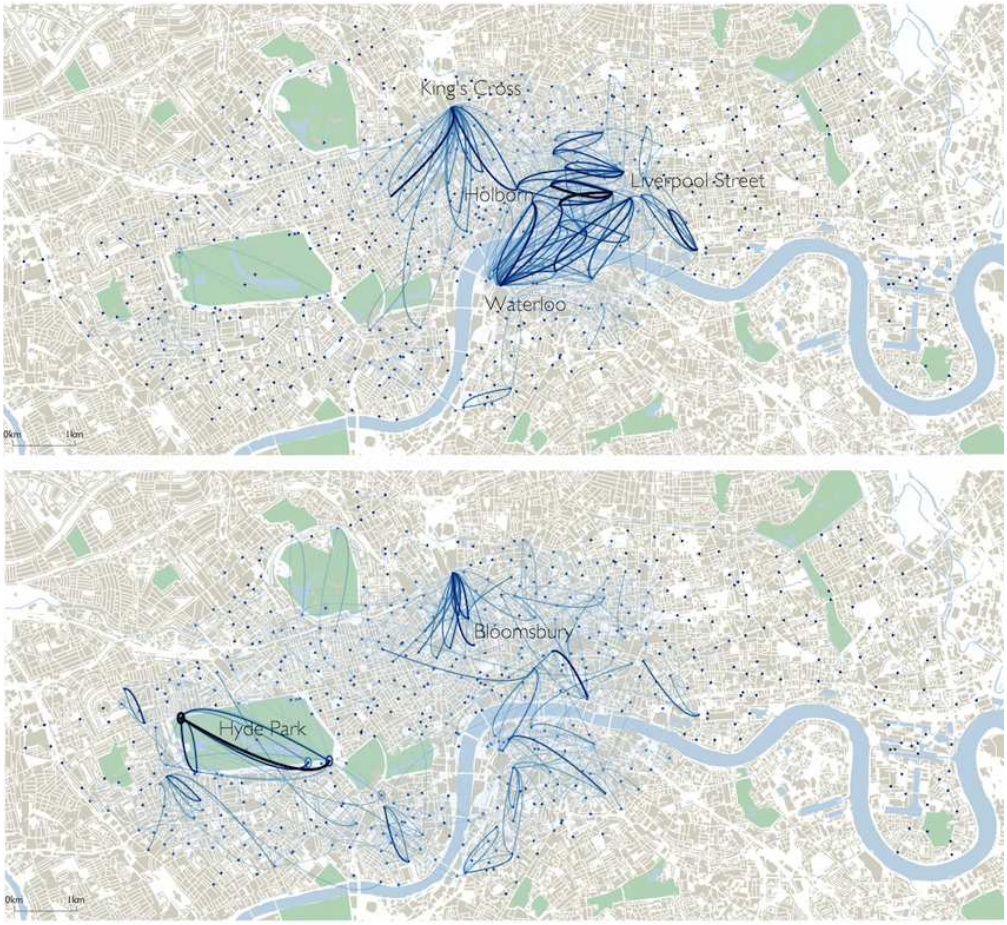

Figure 4.1: Above: journeys between London's major commuting rail terminals – King's Cross, Waterloo and Liverpool Street – are visually salient when all journeys made by men are selected. Below: journeys made by women are selected, with trips within Hyde Park particularly dominant. All journeys made from the scheme's inception through to 14th September 2011 are shown. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

4.3.1 Comparing all journeys and members

First, the full population of members using the scheme from its inception through to 14th September 2012 is studied. An initial observation is that women are under-represented amongst LCHS members, representing just a quarter of the LCHS user population. They registered with the scheme at similar times to men. After a significant surge in interest at the scheme's launch in July 2010, there were more modest increases in demand amongst

both male and female cyclists during January and Summer 2011 and other slight increases in registrations in early Spring and July 2012. Considering the derived variables introduced in Chapter 3, there are substantial differences in both the geodemographic and behavioural profile of male and female LCHS cyclists. There are higher proportions of women apparently living in urban communities than would be expected given the member population as a whole and much fewer in affluent, semi-rural communities. Female members also appear to be far less active scheme users than men. Whilst they comprise 26% of all LCHS members, women make up only 17% of members within the top RF segment - of heavy and recent scheme users - and 34% of members in the bottom RF segment.

Querying journeys made by men and women within the visual analysis application described in the previous chapter, these differences in usage characteristic appear to have a distinct spatial and temporal expression. For men, flows between London's major rail terminals and workplaces – between Waterloo, Liverpool Street, central London and the City of London (Figure 4.1) – overwhelmingly dominate the map view and there are higher than expected flows during weekdays, coinciding with commuting peaks. By contrast, for female members, journeys within London's parks and round trips - those that finish at the same station they started at - dominate. Weekend journeys also constitute a much larger share of all journeys made by female cyclists: 22% of trips made by women take place at weekends, whilst for men this figure is just 16%. The effect size (ϕ_c) for this difference in relative numbers of weekend journeys is 0.1, RR 1.4.

After exploring these data within the visual analysis application, particularly within the first three months after the scheme's launch, an observation was made that retention rates are particularly poor for women. Many female cyclists are within a group who, living relatively close to the scheme's boundary, registered with the LCHS when it first launched, but after experimenting with the scheme by making a small number of 'leisure' journeys ostensibly within London's Hyde Park (Figure 4.1), decided not to use it on a regular basis. The travel behaviours identified for this group of early 'detractors' resonates with the anecdotal and high-level analysis carried out by policy-makers at TfL. Partly due to the LCHS's high profile at its inception, the first three months of usage they regard as atypical. In order to better understand more established behaviours, only journeys over a 12 month period are selected: from 14th September 2011 to 14th September 2012. This serves as the main period of study for the analysis chapters that follow. It amounts to over 5 million journeys made by more than 80,000 members.

4.3.2 Comparing September 2011-2012 journeys

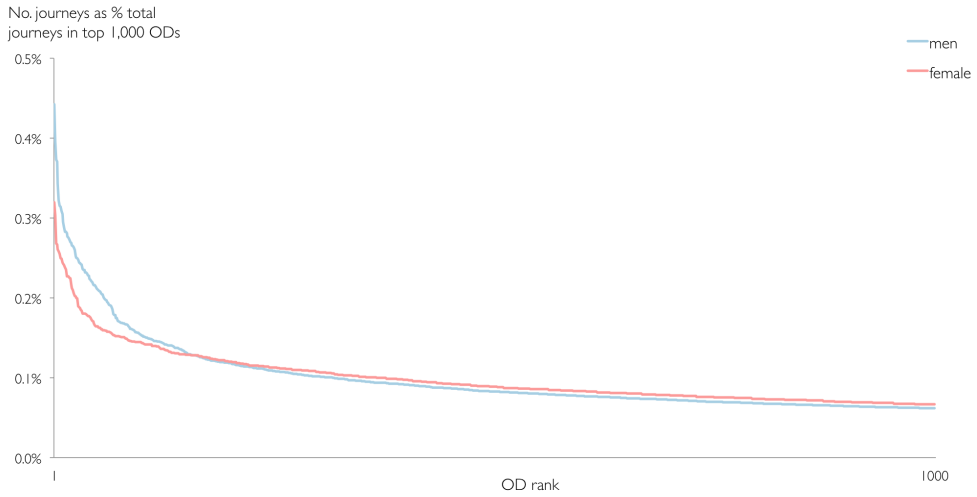

Figure 4.2: Rank-size distribution of 1,000 most commonly made journeys for male and female customers.

Analysing all journeys made between September 2011-2012, then, many of the previously identified differences are preserved. There are fewer female cyclists in the highest RF group than would be expected given the member population as a whole and women are over-represented amongst the lowest RF scores. The relative number of weekend journeys is greater for women than it is for men and there are fewer than expected women amongst the faster travel-time $z - scores$. Exploring journeys within the visual analysis application, men's cycling behaviours again remain highly regular: journeys between major rail terminals and the City of London are clearly visible. For women, however, cycling behaviours are more varied. Journeys within Hyde Park and west London are visually salient, but journeys within parts of central London now become more visible.

One means of quantitatively testing the prominence of commuter flows made by men is to calculate the total share of journeys involving hub stations. Hubs are generally large docking stations located at two major rail terminals – King's Cross and Waterloo – and at the intersection of the City of London and central London (Holborn, labelled in Figure 4.1). In order to cope with very high demand at peak times, bikes are continually replenished at, or withdrawn from, these strategically important stations. Compared with women, substantially more men make journeys that either start or end at a hub station: 31% of men versus 21% of women (ϕ_c 0.1, RR 1.5).

A more detailed study of the most common journeys made by men and women appears in Figure 4.2 and 4.3. In Figure 4.2, journeys (OD pairs) between specific docking stations are ranked according to their frequency. As the figure shows, plotting these ranks and sizes reveals a power-law distribution (Reed 2001) whereby rank position is inversely related to journey frequency. Whilst both curves for men and women follow this familiar distribution, the gradient on the curve is slightly steeper for women, suggesting that the rank-size effect is severe.

Studying these heavily repeated journeys within the visual analysis application, they can be explored in greater detail and inferences made as to their context and purpose. Figure 4.3 shows the 100 most common journeys made by male (top) and female (bottom) cyclists. For men, there is a familiar spatial and temporal pattern, with journeys almost exclusively suggesting a commuter function: weekday journeys between 6am-9am and 4pm-7pm account for 75% of all journeys (this figure for women is 62%), with weekends only accounting for 3% of these journeys. When analysing women's top 100 journeys, a large number also coincide with weekday commuting times. This might be expected since these are heavily repeated trips. Notice though, that it is only the morning peaks that are over-represented. Inspecting all journeys made by 'commuting' female members – those within the high RF segments – this pattern is reinforced: when commuting, female LCHS cyclists are more likely than men to make journeys in the morning peak. Unlike the patterns observed for men, though, weekend journeys are not entirely absent. Around 11% of the top 100 journey combinations for women are made at weekends and, inspecting the map view, 'leisure' journeys within Hyde Park remain visually salient. A number of apparently utilitarian journeys between King's Cross and the Bloomsbury area of London (highlighted in Figure 4.3) can also be seen: 19% of women's top 100 journeys are made within this area, whereas for men this figure is 8% (ϕ_c 0.2, RR 2.5).

There is a sense here that, even when making apparently utilitarian journeys, female members may preferentially select more cycle-friendly parts of the city. Journeys between docking stations at either side of the River Thames - journeys that generally involve relatively large, multi-lane roads and busy junctions - are rare. Instead at peak times, journeys around the Bloomsbury area (Figure 4.3), where roads are narrower, a number of traffic calming measures have been introduced and cycle lanes are slightly offset from major roads, are more common. It is possible to further quantitatively test the finding that women make fewer journeys that involve a river crossing by filtering only those journeys. Whilst 48% of men have made journeys that involve a river crossing, this figure for women is 39% (ϕ_c 0.1, RR 1.2).

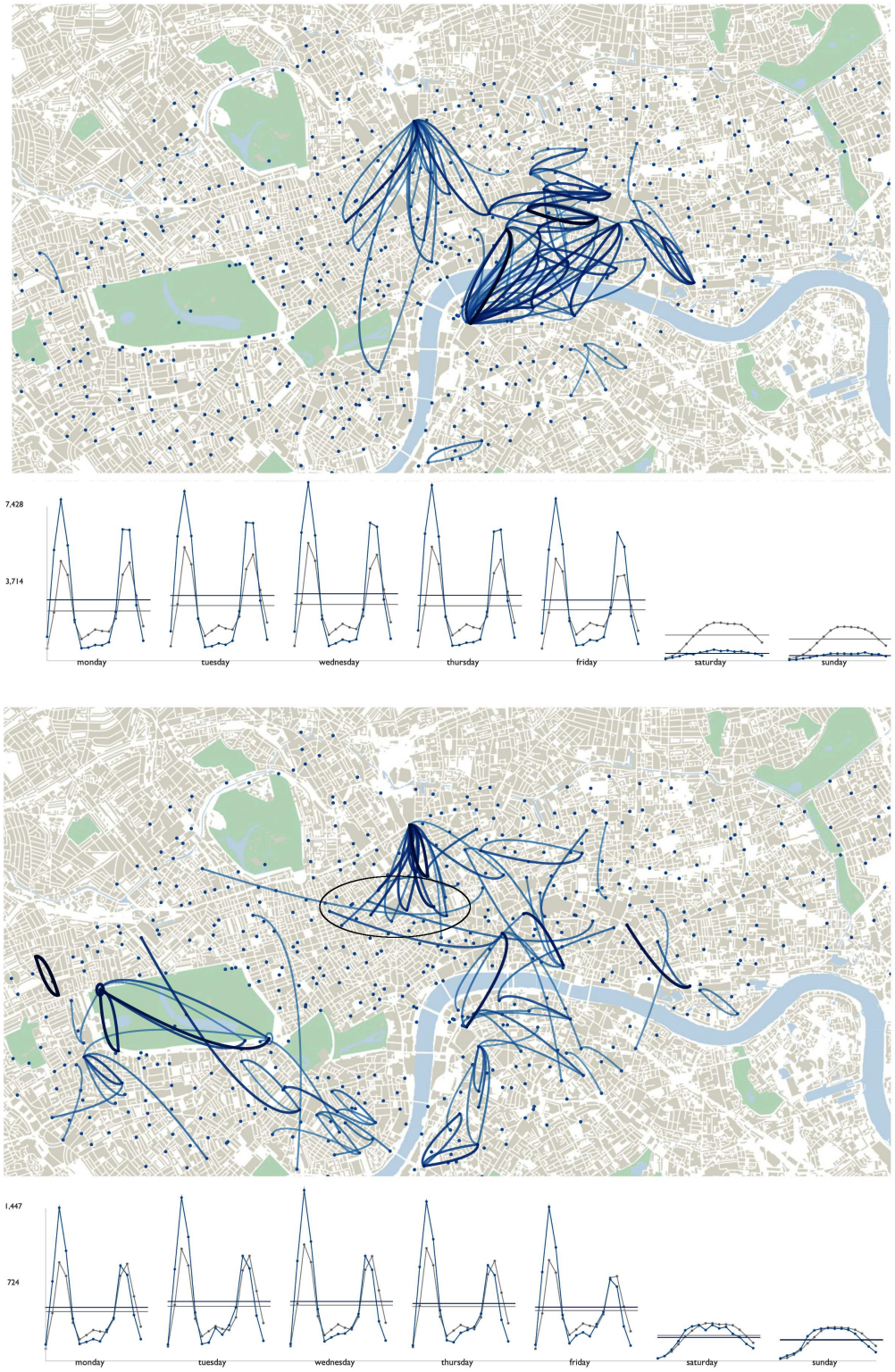

Figure 4.3: Top 100 journey pairs made between September 2011-2012 by male (top) and female (bottom) cyclists. Docking stations within Bloomsbury area are highlighted. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

4.3.3 Controlling for geodemographics

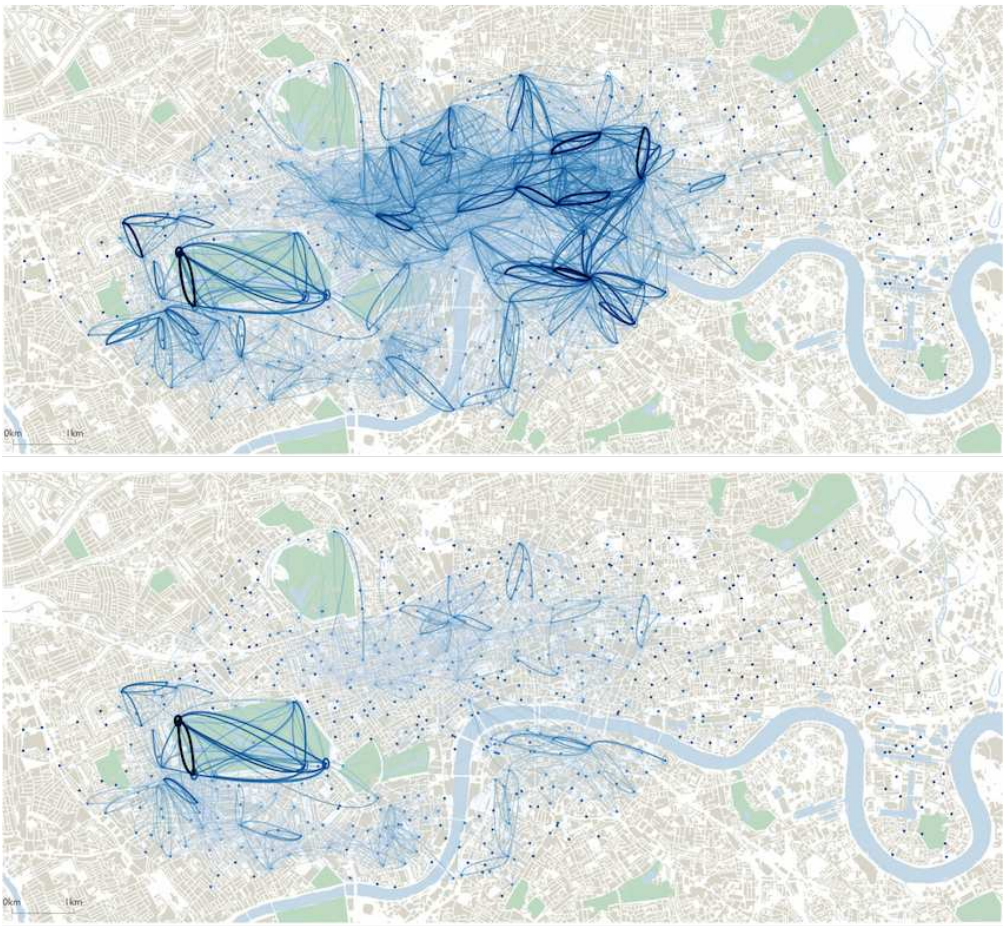

Figure 4.4: Journeys made by male (top) and female (bottom) bikeshare cyclists living <5km from a docking station during interpeak times (10am-4pm Monday - Friday). Flows are coloured by the number of unique members making them. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

Although these gendered differences are true of the total member population, they should not be treated as essential. The dominant pattern when querying male users is of commuter travel. Highly visible amongst these journeys is a group of users typically living in semi-rural and suburban communities (the OAC classification Suburbanites) who, after commuting into London on a train, routinely use the scheme to make highly regular journeys from the major rail hubs identified at the start of the chapter. Women are under-represented amongst the non-London member population. Whilst 28% of men subscribing to the scheme live more than 15km from a docking station, this figure for

female cyclists is just 16% (ϕ_c 0.1, RR 1.8). Women are therefore under-represented amongst this group of very heavy scheme users and in directly comparing male and female cyclists, judgements are made about two very different populations. It is possible to control for these differences within the visual analysis application by selecting only members who apparently live less than 5km from a docking station. This subset represents over 50,000 members and almost 3.3 millions journeys made between September 2011-2012.

Immediately there is greater convergence between male and female cyclists. Though to a lesser extent than for women, men living this distance from a docking station become slightly over-represented amongst weekend journeys. The spatial patterns of men's journeys are now far less regular, with journeys between Waterloo and the City of London no longer dominating the map view. Flows within Hyde Park and west London can also be identified and journeys within the eastern expansion area (opened in March 2012) are now visible. A very diverse set of journeys is also found extending into the semi-residential areas of the east and south east of the city.

Although the spatial and temporal pattern of journeys made by men has changed, a number of the differences previously identified remain. Women are under-represented amongst the high RF scores, amongst the faster travel time $z - scores$ and for those in the top four RF groups, amongst the morning rather than evening peaks. Women are also slightly less likely to make journeys interpeak: journeys taken on weekdays between 10am–4pm are less common for this subset of female members. In addition, when selecting on these interpeak journeys, there are substantial differences between the types of journeys being made.

These spatial differences are again best expressed when evaluating the spatial view, but altering the way flows are emphasised. A problem with weighting each unique OD pair according to the absolute numbers of journeys being made is that when filtering by multiple variables, the sample size obviously reduces. For example, by filtering on gender, distance from docking station and by time of travel (interpeak weekday journeys), only 136,000 journeys made by 10,000 women and 500,000 journeys made by 29,000 men are considered. Whilst this remains a very large amount of data, it is occasionally the case that a minority of members repeatedly making the same journey (OD) pair in this time period can appear overly salient when weighting flow lines by absolute journey frequency. Weighting flow lines instead according to the number of unique members making those journeys, partially overcomes this problem. Figure 4.4, then, emphasises where most

members living less than 5km from a docking station make interpeak journeys. Whilst large numbers of men do appear to make interpeak journeys within Hyde Park, a diverse set of journeys in other parts of the city are found. Some of these flows suggest leisure cycles, with various journeys from east to west along the popular south side of the river – the Southbank area – easily identifiable. Others appear more utilitarian in nature, with many short journeys made within central and the City of London. For female cyclists, spatial travel behaviours appear to be more constrained: most tend to make interpeak journeys within Hyde Park and west London and comparatively few might be regarded as utilitarian.

Since the spatial travel behaviours of male and female LCHS cyclists are so different, a behavioural control is finally added: only members living less than 5 km from a docking station and who are in the top four RF groups – the most active and experienced LCHS users – are selected. The spatial views for this subset of men and women appear in Figure 4.5. Again, this sample of nevertheless frequent female scheme users tend to select more particular parts of the city than do men. Journeys within west London and Hyde Park dominate the map view and outside of this area journeys are relatively spatially constrained. Whilst large numbers of male cyclists do make journeys within Hyde Park, there remains a relatively diverse set of journeys across central London, the City of London and further east (Figure 4.5). Remember, only the most active and experienced scheme users, who live relatively close to a bikeshare docking station, are compared.

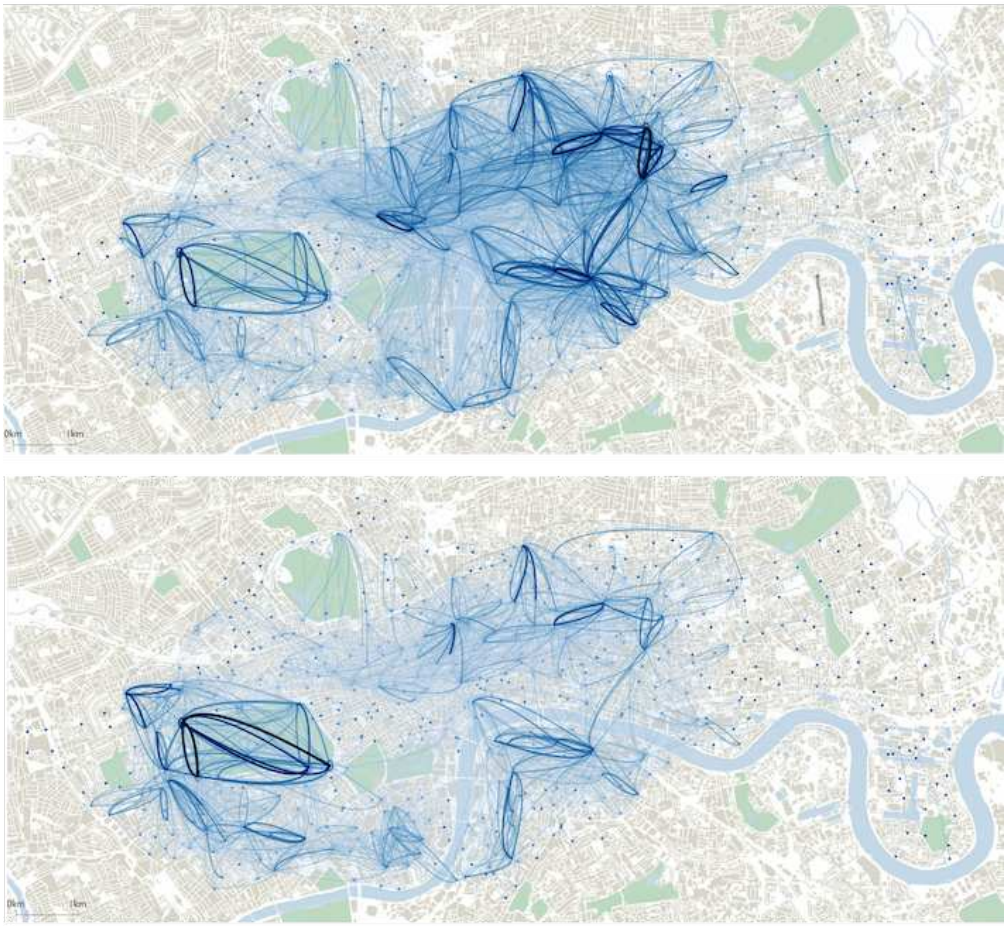

Figure 4.5: High RF men and women living <5km from a docking station. Flows are coloured by the number of unique members making them. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

4.4 Discussion

Exploring the LCHS dataset using the designed visual analysis application, distinct cycling behaviours were in this chapter quickly identified, located within their spatiotemporal context and related to the personal characteristics of the members making them. The findings from this analysis resonate with the already substantial research on gender and urban cycling behaviour discussed in Section 4.1.

The first main finding was that scheme usage amongst men is highly regular, suggesting a strong commuter function, whilst leisure orientated journeys appear to be more dominant

for female members. Although some of this variation can be explained by the different geodemographic characteristics of the member population, that apparent commuting journeys are far less common for women is instructive. Elsewhere, survey based studies into claimed cycle behaviours have found women are generally less likely than men to cycle for commuting purposes (Heesch et al. 2012). This has also been confirmed by detailed analysis of observed behaviours, albeit based on a much smaller dataset (Dill & Gliebe 2008).

Secondly, even after controlling for variations in the geodemographic and behavioural characteristics of LCHS members, many important differences were identified. Women are consistently over-represented amongst the least heavy LCHS users; they are routinely under-represented amongst the faster travel time $z - scores$; and the temporal structure of women's journeys suggests slightly greater levels of cycling at weekends. These findings again appear consistent with related research, which finds that women generally cycle at slower speeds than men (Dill & Gliebe 2008).

A more substantial insight was that women's journeys appear to be highly spatially structured. Since female members are more likely to make weekend journeys, it is not surprising that west London and Hyde Park, relatively leafy parts of the city, dominate when querying their journeys. However, there is a sense that women preferentially select very particular parts of the city, even when making apparent utilitarian trips. Journeys around the Bloomsbury area, where roads are narrower, a number of traffic calming measures have been introduced and cycle lanes slightly offset from major roads, are amongst the most common journeys made by women at peak times. These findings also appear consistent with existing research. Various survey (Tilahun et al. 2007) and observation-based studies (Dill & Gliebe 2008), including a census of cyclists (Garrard et al. 2008), have found a preference amongst female cyclists for low traffic streets, with routes offering the maximum separation from motorised traffic found to be a particularly high priority.

Two new discoveries can also be offered from the analysis presented in this chapter: that female members who use the scheme at peak commuting times are more likely than men to make those peaktime journeys in the morning than the evening peaks; and that female members are less likely than men to make apparent utilitarian journeys during the working day, irrespective of how often they use the scheme. Clearly, these findings may be particular to bikeshare scheme usage rather than more general cycling behaviour. Their underlying motivations may nevertheless reflect more fundamental differences in

attitudes towards, and perceptions of, cycling.

Finally, it is worth reflecting on the detail and scale of analysis achieved in this exploratory analysis chapter given this study's research objectives and questions (see Chapter 1). Perhaps the most interesting aspects of the discussed analysis relate to the spatiotemporal context under which men and women make journeys: for example, the relative use of commuter hub docking stations at peak times, or use of docking stations within parks at weekends. These spatiotemporal aspects of behaviour, along with important geodemographic controls, can be reliably investigated because the LCHS provides information on a full population of users and their journeys. Whilst the literature summarised in Section 4.1 is large and diverse, the scale and scope of observational based studies – studies that consider 'real' rather than claimed behaviours – is relatively limited. Dill & Gliebe's (2008) GPS-based study is one of the most comprehensive of its kind, but achieved a sample of just 166 participants. Analysis of detailed travel behaviours at the scale of the city, as appears in this chapter, would clearly not be possible with the much smaller, actively-collected datasets. The depth of insights into spatiotemporal cycle behaviours generated from this exploratory analysis, then, suggests that distinct cycling behaviours can be identified from the LCHS usage data (RQ1). That the identified behaviours relate strongly to existing research, suggests that *individual* LCHS cycling behaviours are meaningful and potentially generalisable – a concern discussed in Chapter 2 – and that the findings may represent useful contributions to the Transport Studies domain (Overall RQ).

4.5 Moving forward

By exploring the spatiotemporal and customer-related context under which journeys are made, several different types of usage behaviour were identified in this chapter. An important aspect of behaviour is that of commuting. In the following chapter, an approach is proposed for identifying and labelling commuting journeys with greater certainty; and which involves a spatial analysis of individuals' scheme usage. As discussed, that the LCHS dataset provides a complete and spatially and temporally precise record of behaviour is significant and the commuter classification would not be possible if only a partial and less precise set of information on individuals' journeys was recorded. As a result of the commuter classification, commuting behaviour can be queried and described with greater certainty. The commuter classification also provides an important

additional contextual variable – a spatial reference point for the likely workplaces of commuting members. This information is used to question early explanations suggested in this chapter around the observed spatial differences in travel behaviours between male and female cyclists.

Chapter 5

Labelling and studying commuting

Abstract

In this chapter, a technique for automatically classifying commuting behaviour is developed that involves a spatial analysis of cyclists' journeys. A subset of potential LCHS commuting cyclists is identified and for each individual, a plausible geographic area representing their workplace is defined. All peaktime journeys terminating within the vicinity of this derived workplace in the morning, and originating from this derived workplace in the evening, are labelled commutes. Three techniques for creating these workplace areas are compared: a weighted *mean-centres* calculation, spatial *k-means* clustering and a *kernel density-estimation* method. Visual analytics software is designed and used to support this evaluation. Exploring the three proposed techniques within this software, commuters' peaktime journeys are found to be more spatially diverse than might be expected and for a significant portion of commuters, there appears to be more than one plausible spatial workplace area. The density-estimation is selected as the preferred method for labelling commuting behaviour. Once all commuting events are labelled, two distinct types of commuting activity are identified: those taken by LCHS customers living outside of London, who make highly regular commuting journeys at London's major rail hubs; and more varied commuting behaviours by those living very close to a bikeshare docking station. Interpeak journeys apparently taken as part of cyclists' working day are found around London's universities. Exploring identified workplaces within a further set of software, imbalances in the number of morning commutes to, and evening commutes from, derived workplaces are identified and differences in the geography of men's and women's workplaces are also discovered that appear to relate to the actual geography of employment in the city. This latter observation is important, as it illustrates some of the challenges behind offering explanations for observed spatial travel behaviours when using such passively-collected data.

This work has been published in: Beecham, R., Wood, J. & Bowerman, A. (2014) Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, 47(September), pp.5–15.

5.1 Research context

Early analysis of the LCHS dataset revealed very heavy scheme usage during peak times (Wood et al. 2011, Lathia et al. 2012) and the exploratory analysis in the previous chapter certainly suggests a strong commuter function amongst particularly male bikeshare cyclists. However, the LCHS is clearly also used for purposes other than commuting; and this might be true even of journeys made during the weekday peaks. This chapter aims to label commuter journeys and commuting members with greater certainty, thereby allowing the commuting behaviours suggested in Chapter 4 to be investigated in a more formal way. Its main focus is analytical; the overriding research question asks:

- **Analytic question:** How can commuting journeys and commuting LCHS cyclists be reasonably detected?

The task of identifying commuting behaviour might initially seem like a straightforward data mining exercise. One means of identifying commuting journeys might be to find all instances where a LCHS cyclist completes a closed peaktime 'loop', where their last journey of the day happens during the evening commute and is the inverse of their first journey of the day. Recent analysis of usage data from London's metro system has found, however, that such assumptions about commuting behaviour often do not hold (Lathia et al. 2013). With a month's usage data from the London Underground, Lathia et al. (2013) aimed to identify the nature and extent of commuting behaviour. They made reasonable assumptions about commuting travel behaviours: that commuters will make on average two journeys or more per day, that they will typically repeat the same origin-destination (OD) pair and that commuters' behaviours will form a closed loop whereby the first origin and last destination of the day are the same. The authors subsequently found, however, that many travellers did not fit these expected patterns. Sixty-six percent of users took less than one trip every two days and only 8% met the

expected two trips per weekday criteria (Lathia et al. 2013). The authors later proposed clustering algorithms for automatically finding groups of travellers with similar temporal travel profiles – who typically travel at particular times of day and days of the week.

The approach to temporal clustering taken by Lathia et al. (2013), and also previously by Agard et al. (2011) using bus transit data, is very similar to the behavioural clustering described in Chapter 3. It enables those who apparently use a transport system almost exclusively for commuting (in this research *9-to-5-ers*) to be distinguished from those with other usage characteristics. However, identifying these people alone does not give a total, journey-level view of commuting. As the large *anytime* user group introduced in Chapter 3 suggests, it is reasonable to assume that those who commute may also often use the LCHS for non-commuting, leisure-oriented or utilitarian weekend journeys. If commuting users are defined only as people whose dominant travel patterns coincide with commuting times, then it is possible to miss the commuting behaviour of individuals who typically use the scheme for other purposes.

An alternative approach, taken in this chapter, is a spatial analysis of LCHS users' peaktime journeys. A broad *spatial* area representing each cyclist's workplace is identified and all peaktime journeys arriving at this workplace area in the morning, and departing from this workplace area in the evening, are labelled as commutes. Clearly there is no record of such workplace areas in the customer database; and instead they are derived from exploring spatial patterns of LCHS cyclists' peaktime journeys.

Once commuting events are detected using this method, three thematic questions are posed:

1. What are the characteristics of people who take part in commuting based activities?
2. Where do commuting events happen?
3. Under what circumstances are journeys made during the working day?

The chapter concludes by reflecting on the commuter classification given the research questions outlined in the Introduction chapter.

5.2 Use of *visual analytics*

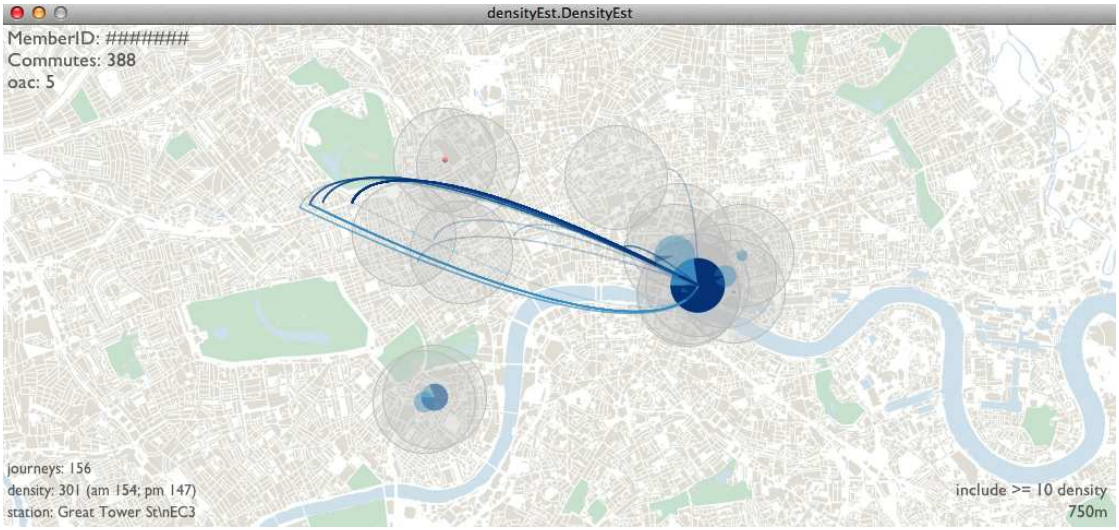

Figure 5.1: Example *visual analytics* systems used later in the chapter. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

A substantial part of this chapter involves the development and evaluation of spatial analysis techniques for deriving cyclists' likely workplaces. Several commonly used spatial data mining techniques are described and proposed. With relatively few studies tackling similar problems using similar datasets, however, decisions must be made about the suitability of these techniques and their underlying parameters. A set of *visual analytics* applications (see Figure 5.1) are built for making these decisions and the use of visual techniques in this chapter should be regarded slightly differently to the exploratory visual analysis that appears in Chapters 4 and 6.

Visual analytics is a sub-discipline or extension of Information Visualization that is concerned with the processes and decisions made by an analyst when struggling with creating insights from raw datasets (Keim et al. 2010). For Keim et al. (2010), an analyst might start by computing various value-free models – models that are generally not informed by hypotheses. The analyst then reviews and refines these models by interacting with the dataset visually, using interactive graphics: he or she may modify certain parameters, select other analysis algorithms and/or spatial and temporal units. In doing so, the analyst is better able to develop and organise schemas – mental models about a concept or phenomenon that are formed through both empirical information and the analysts' experiences (Priolli & Card 2005). *Visual analytics* therefore considers how interactive

graphics can support analytical decision-making: it accepts that human and subjective decisions underpin the data analysis process (Thomas & Cook 2006).

Although the techniques proposed for identifying customers' workplaces cannot be described as value free – underlying them are assumptions about how LCHS cyclists commute – it is certainly the case that subjective decisions are made about their suitability. These subjective decisions are supported through the use of interactive graphics and the new analysis applications introduced in this chapter, and designed specifically for this purpose, are therefore here referred to as *visual analytics* systems.

5.3 Data processing

5.3.1 Labelling commuting events

Potential commuters

The commuter classification aims to identify those journeys that regularly end within the vicinity of a customer's workplace in the morning and regularly start from within the vicinity of a customer's workplace in the evening. For convenience, such locations are labelled 'workplaces'; however, there may clearly be other attractors for repeated peaktime activity. In order to label individuals' workplaces with any confidence, it is first necessary to find a reasonable number of data points for each member using the scheme at peak times.

One means of identifying this population of 'potential commuters' may be to filter all journeys taken during the morning (6am-10am) and evening (4pm – 8pm) peaks and for each customer, identify the total number of these peaktime journeys. Clearly some cyclists, particularly shift and casual workers, will also make commuting journeys outside of peak times. However, by restricting to repeated journeys within the working day, there is perhaps greater certainty of identifying genuine work-related activity. Ordering customers according to the number of peaktime journeys they make and plotting these frequencies (Figure 5.2), reveals a heavy-tailed distribution: a large portion of members making peaktime journeys (45%) make only between 1 and 10 such journeys. There is, though, no obvious break in this distribution that suggests more emphatically a number of repeated peaktime journeys beyond which a member is likely to use the scheme for commuting purposes.

Exploring the spatial and temporal context of individuals' journeys visually better enables judgements about the nature of cycling behaviours to be made. Using the visual analysis application introduced in Chapter 3, the space-time structure of journeys made by cyclists who make few peaktime journeys are compared with those who make many peaktime journeys. After approximately 20 trips, customers' usage characteristics begin to increasingly suggest a commuter function. This is expressed in Figure 5.3, which shows all journeys taken by the selected subset of members (those making between 21 and 30 journeys at peak times). Inspecting the map view, reciprocal journeys between the main commuter docking stations identified in the previous chapter become increasingly visible

for those making between 21-30 peaktime journeys. By contrast, selecting individuals making less than 20 repeated peaktime journeys, the observed temporal and spatial pattern suggests a leisure function, with largely weekend journeys within London's parks more prominent.

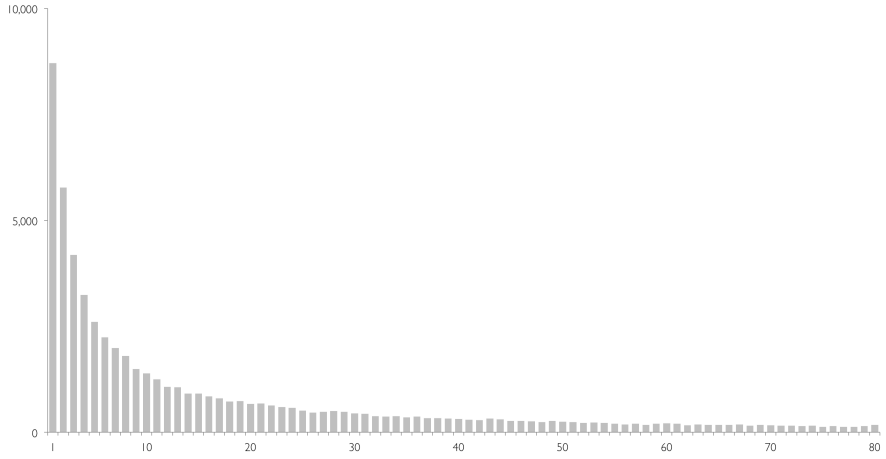

Figure 5.2: Frequency distribution of peaktime journeys taken by LCHS customers.

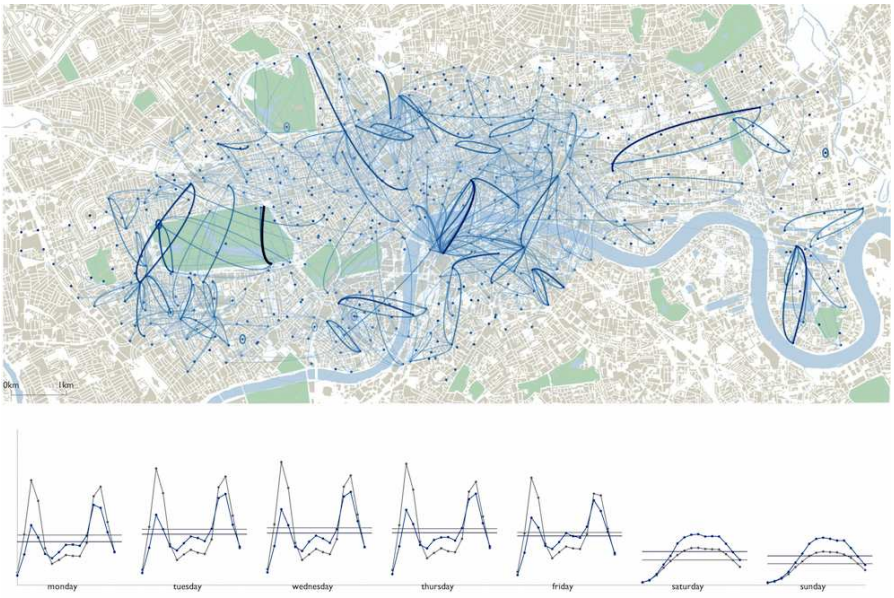

Figure 5.3: Journeys made by LCHS customers completing between 21-30 peaktime journeys are shown. For members making at least 20 repeated peaktime journeys, travel behaviours begin to suggest a commuter function: 39% of cyclists making between 21-30 peaktime journeys have completed at least one journey involving a commuting hub station; for those making between 1-10 peaktime journeys this figure is noticeably smaller, at 19%. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

Mean-centres method

Figure 5.4: Application for validating commuting mean-centres. Left: tri-centric distribution of workplace locations for a single member. Right: bi-centric distribution of workplace locations for a single member. Standard deviation ellipses are drawn in grey; blue dots represent commuter destinations (am) and origins (pm) and are sized by relative frequency. A suggested 500m buffer for excluding journeys appears in red. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

In total, 34% of the scheme's member population are potential commuters; making more than 20 repeated peaktime journeys in the 12 months between 14th September 2011 and 14th September 2012. For each cyclist in this group, the aim is to derive a spatial area that represents their workplace. First, a very simple technique for deriving these workplace centres is explored. After identifying all peaktime destinations during the morning and origins during the evening, the (frequency) weighted centroid of docking station locations each cyclist uses during peak times is calculated. An assumption is made that all inbound journeys in the morning, and outbound journeys in the evening, that lie within a user-defined distance – an acceptable walking distance – of this mean-centre might be labelled commutes.

A visual analytics system is developed to evaluate the mean-centres classification. This software enables spatial patterns of individual cyclists' peaktime journeys to be identified, along with the (frequency) weighted centre of these locations and a buffer around this centre for including commuting journeys. A screenshot from this software appears in Figure 5.4. Blue dots represent docking stations (potential workplace locations) and are sized according to the number of peaktime journeys that customer makes to/from those docking stations. In order to evaluate the spatial dispersion of these locations, a standard deviation ellipse (Yuill 2011) appears in grey, the weighted-centre of which represents that customer's hypothesised workplace. In red, a buffer for filtering journeys is drawn. Through a simple set of key presses, it is possible to distinguish between

journeys made during the morning and evening peaks; to adjust the spatial buffer drawn around each centroid; and to iterate through and evaluate the classification for each member. It is also possible to order customers according to the spatial dispersion values calculated when producing the standard ellipse; and therefore to identify how successful the algorithm is when customers do not have a spatially coherent set of workplaces.

Iterating through the customer database, the initial 'walking distance from centre' threshold of 500 metres (m) appears to be too conservative: it unnecessarily excludes journeys that might be reasonably linked to customers' mean-centres – customers' potential workplaces. In many instances, there are spatial outliers that substantially influence the workplace classification. Here, journeys to docking stations a distance away from individuals' main workplace centres have the effect of displacing cyclists' weighted mean-centres. Perhaps more importantly, a large number of members have more than one workplace (Figure 5.4): the distribution of points is either bi- or occasionally multi- centric. In these instances, simply using the weighted mean-centre of all peaktime commuting points would serve to exclude almost all journeys an individual makes which, since they remain spatially clustered, will likely represent genuine commuting journeys. Without representing individuals' peaktime journeys visually, these problems would perhaps not be as immediately obvious. Attending to the standard deviation of cyclists' potential workplace locations, and ratios of the maximum and minimum dispersion calculated when creating the standard deviation ellipses, it is estimated that 20% of commuting LCHS users are likely to make journeys to or from more than one spatial workplace cluster.

k-means method

A second proposed solution, which might overcome at least one shortcoming of the mean-centres method (the fact there are multiple workplace clusters), is the *k-means* clustering algorithm. The *k-means* algorithm is run on each cyclist's set of potential workplace docking stations, using information from the standard deviation ellipses to specify the number of workplace clusters (k) a cyclist is likely to have. If the ratio of dispersion and standard distance exceeds a certain threshold, then k=2 or k=3 cluster centres are specified. Alternative methods, such as *hierarchical cluster analysis* (HCA) (O'Sullivan & Unwin 2002), might better enable an appropriate number of clusters to be first suggested. However, the output of each HCA would need to be inspected at an individual cyclist level and scaling this analysis to the full LCHS customer population would therefore be problematic.

Running *k-means* clustering algorithms on a sample of member records, and inspecting the classification within the same set of visual analytics software described in 5.4, it appears that the rules for predicting the number of clusters are not always successful: a high level of distance deviation and large ratio of dispersion does not solely suggest a bi-centric distribution (e.g. Figure 5.5). Separate to this issue of correctly specifying an appropriate number of clusters, the same problem of extreme spatial outliers displacing cluster centres also exists.

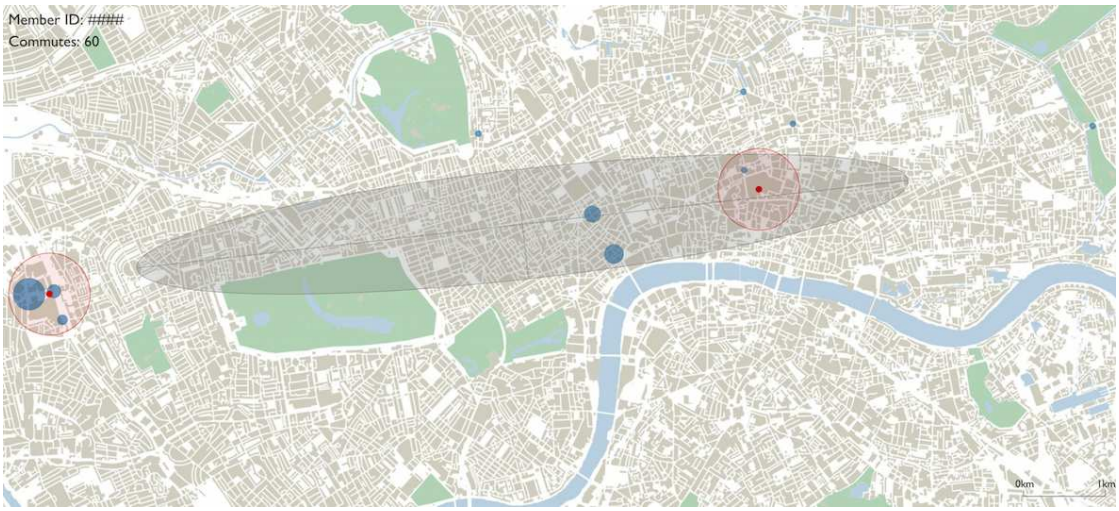

Figure 5.5: *K-means* analysis assuming standard distance and ratio of dispersion criteria are such that a 2-cluster solution is required and therefore k=2. A primary workplace cluster is successfully identified (far left), but a second centre of activity is entirely missed. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

Density-estimation method

Conscious of these two pitfalls, a density-based method (O'Sullivan & Unwin 2002) is proposed for successfully coping with multiple spatial clusters of workplace centres and, importantly, negating the displacement problem caused by spatial outliers. In the proposed density-estimation method, observations are made at every potential workplace docking station. For each LCHS member, all peaktime destination (am) and origin (pm) docking stations are found and, at each docking station, the total number of journeys made to and from that docking station, and to and from neighbouring docking stations that are within a user-defined walking distance, is recorded. This user-defined distance serves as a kernel in the density estimation. Once the density estimates have been made, all journeys to docking station locations whose density counts are above a certain thresh-

old, and therefore whose spatial areas are visited frequently at peak times, are labelled as commuting journeys.

For the density-estimation technique to be successful, decisions need to be made about two parameters: the size of the kernel and the density threshold used to exclude or include peaktime journeys. Again, showing this analysis visually better enables the technique to be evaluated and Figure 5.6 displays an extended set of visual analytics software designed for this evaluation. Docking stations are represented as dots, sized according to the number of peaktime journeys they receive. The kernels used in the density-estimation appear as transparent circles centred at each docking station. They can be resized, and the classification subsequently recalculated, 'on the fly'. Docking stations that meet a user-defined acceptance criteria appear in blue; journeys to and from excluded docking stations (non-commutes) appear in red.

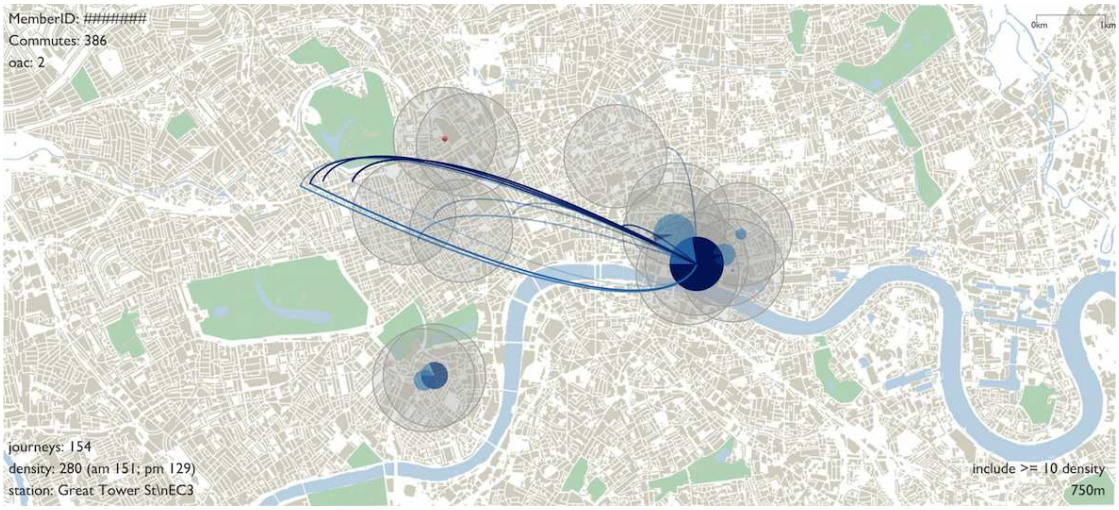

Figure 5.6: Density-estimation for a single cyclist is explored. Docking stations classified as commutes appear in blue; non-commutes in red. Resizable kernels are shown in grey. Morning journeys to workplaces (light blue/red) can be distinguished from evening (dark blue/red) journeys. By mousing over each workplace docking station, morning journeys arriving at, and evening journeys departing from, that docking station are displayed. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

Using this software, it appears that the density-estimation technique begins to address the two previous problems. It is successful at accepting secondary and tertiary clusters whilst still excluding outlier locations. Where members apparently make commuting journeys to more than one spatial workplace cluster, those journeys are included, whilst more exceptional activity is correctly treated as non-commuting behaviour. Since the centres for individuals' workplaces do not need to be calculated, the problem of spatial

outliers displacing workplace centres also no longer exists.

Again, two analytical decisions influence the success of this technique: the size of the kernel and the inclusion density criteria. A radius of 750m around each potential workplace docking station is selected as the preferred kernel width. This kernel size is partly chosen as colleagues at TfL, who contributed to this analysis, reported anecdotal information that 750m is generally an accepted maximum between-docking-station walking distance. By visually scanning the spatial distribution of members' potential commuting docking stations, there is some empirical support to this claim. Iteratively exploring cyclists' peaktime journeys, rarely are there apparent clusters with docking stations extending beyond this 750m buffer. Secondly, a local density threshold of 10 or more journeys is used for excluding docking stations whose total peaktime journey frequency, combined with other docking stations within the 750m kernel, fall below this limit. This is decided by iterating through each individual's journeys, interactively varying the inclusion density parameter and visually scanning accepted and rejected workplace docking stations. With an inclusion density greater than 15, false negatives begin to appear: docking stations that lie within a spatial cluster and should be labelled as commutes, but instead are coloured as red and therefore rejected. An inclusion density of less than five leads to false positives: docking stations visited relatively rarely and that are spatially distinct from an individual's dominant workplace cluster, are coloured blue and therefore wrongly accepted as workplaces.

Using the described visual analytics software, it is therefore possible to quickly identify problems associated with proposed analysis techniques and suggest appropriate threshold parameters. When applied to the full population of LCHS members, there may be instances where a cyclist's commuting behaviour is still labelled incorrectly. However, with no clear rules for quantitatively evaluating each technique, the visual analytics software allows proposed methods to be related directly to real cycling behaviours and, as a result, empirically-derived threshold values to be specified.

5.4 Analysis

5.4.1 Studying commuting behaviour

Setting the density-estimate with a kernel of 750m and an inclusion density of 10 journeys, 34% of the member population making journeys in the 12 month study period are labelled as commuters and in total classified commuting journeys represent 49% of all journeys taken between 14th September 2011 – 14th September 2012.

Who are commuters?

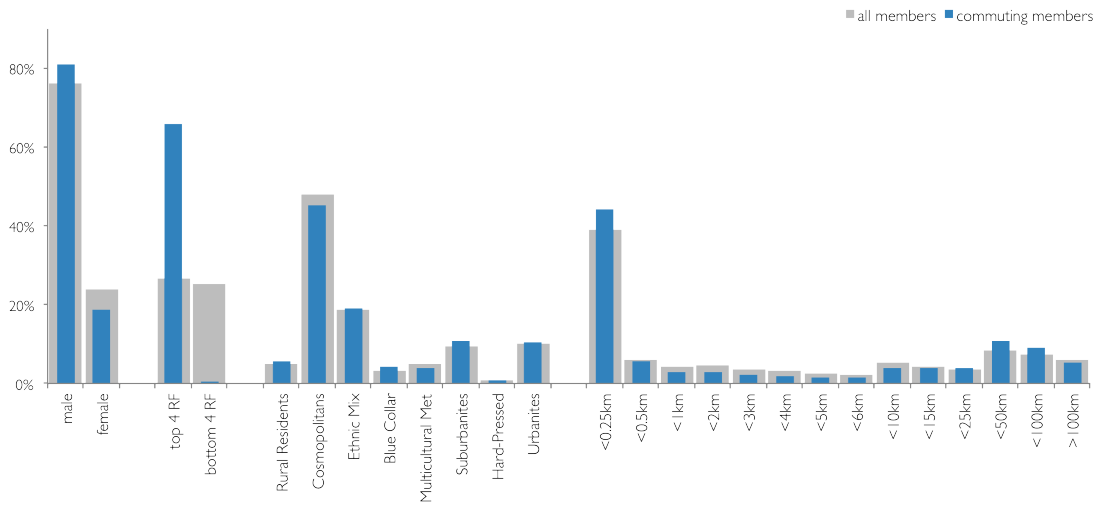

Figure 5.7: Geodemographic profile of commuters.

Figure 5.7 is a geodemographic profile of commuting cyclists identified using the density-based classification. This profile confirms many of the (data-driven) assumptions about commuters that appeared in the previous chapter. Compared to the total LCHS member population, men and high RF users are over-represented amongst commuting members. So too are members who live very close to a bikeshare docking station, but also those that live outside of London (between 15km-100km from a docking station). By contrast, LCHS users that live moderate distances from a docking station (between 500m-10km away) are under-represented amongst commuting members given their share of the total LCHS population.

From this, one might assume that immediate proximity to a docking station is an important motivating factor for commuting usage. Those living less than 500m from a docking station can access their nearest bike station by foot and the same might apply for those living between 15km-100km from a docking station. These individuals may live too far from central London to commute using the London Underground or bus system and instead travel into London using a commuter train which, unlike the bus or underground, arrives only at major rail terminals and therefore possibly some distance from commuters' workplaces. From there, it might be the case that LCHS bikes are accessible and represent a desirable option for completing the final leg of a journey.

Where do commuting journeys take place?

The spatial patterns of journeys made by commuting members seem to support this claim. Commuting journeys tend to coincide with major rail stations and this is particularly true for those who apparently live between 15km-100km from a docking station. For all commuting members living this distance from a docking station, 25% of commuting journeys start or end at a hub docking station located at two of London's major rail terminals – King's Cross and Waterloo. For those living less than 10km from a docking station, this figure is just 5% (ϕ_c 0.3, RR 4.5).

In addition, commuting journeys made by London-resident members appear more spatially and temporally diverse than for those commuting from outside of London. Ordering cyclists according to the proportion of their commuting journeys that are unique, for the median commuter amongst those living less than 10km from a docking station, 32% of commuting journeys are entirely unique OD pairs. This figure for commuters living between 15km-100km from a docking station is just 19%. The implication is that non-London commuters tend to consistently use a reduced number of docking stations when they commute. Additionally, travel times appear to be slightly more balanced between the morning and evening commute for London-resident commuters. For those living less than 10km from a docking station, 54% of commutes take place in the morning peak, whereas for those living 15km-100km from a docking station, this figure is 56%.

Who makes interpeak working-day journeys?

Labelling all commuter journey *events* in the dataset, rather than simply commuting members, allows an interesting question to be asked: to what extent do bikeshare cyclists make journeys as part of their working day, after having commuted into work in the morning and before commuting home from work in the evening? Each interpeak journey (weekdays between 10am-3pm) made by commuting members is considered and where, on the same day that member makes a commuting journey in either of the morning or evening peaks, the original interpeak journey is labelled as an interpeak 'working-day' journey. In total, 78% of commuting members make such interpeak working-day journeys.

There is some concentration of interpeak working-day journeys around London's universities. Docking stations around the Bloomsbury area, where three universities are located, are a focus of interpeak working-day activity and so too are journeys around a major university towards the south west of Hyde Park. Spatially filtering these journeys, the lunchtime peak is less severe in those parts of London with a concentration of universities: 22% of interpeak working-day journeys that involve docking stations within the vicinity of universities are taken between 12pm-1pm, whilst this figure for journeys within the City of London, London's commercial centre, is 26% (ϕ_c 0.1, RR 1.2). One possible explanation for the smoother distribution of interpeak working-day journeys around universities could be delayed commutes taken by individuals employed or studying at universities and with comparatively flexible working hours.

Geography of workplaces

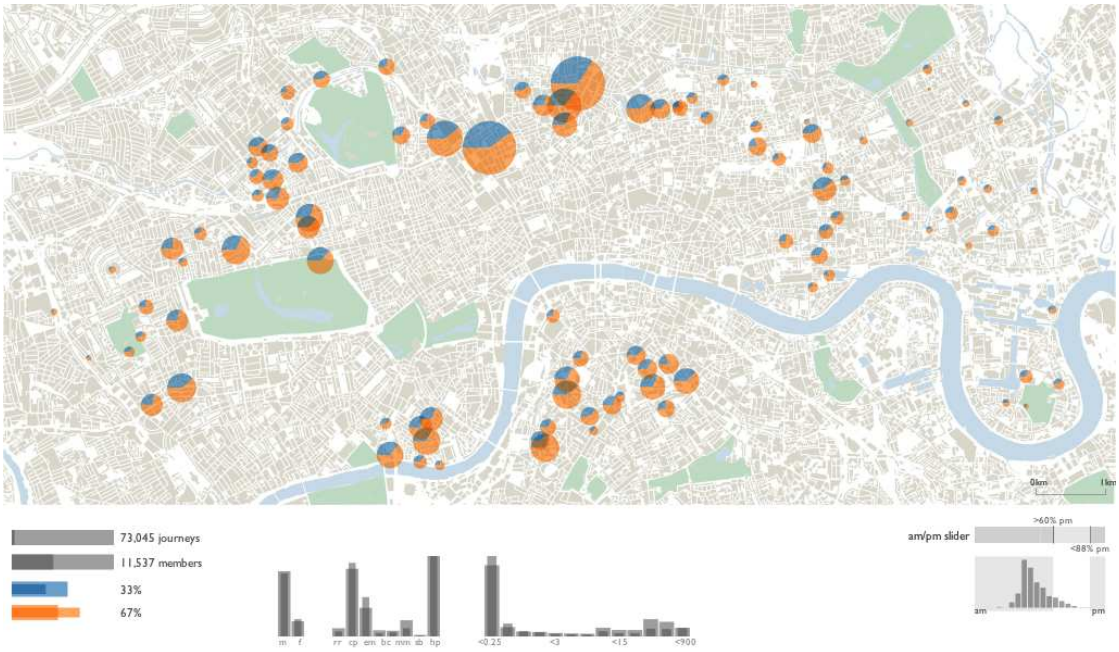

Figure 5.8: Application for exploring 'global workplaces'. Map: pie charts are workplace docking stations sized according to number of commutes arriving (blue) in the morning and departing (orange) in the evening. Bottom: gender and geodemographic variables appear as bars; in am/pm slider, docking stations where more evening commutes depart than morning commutes arrive are selected. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

Section 5.3.1 demonstrated how, through visually depicting techniques for identifying workplace docking stations, problems with each technique were evaluated in the context of individual cyclists' journeys. This analysis process, and especially the design addition whereby morning journeys are distinguished from evening journeys, also reveals interesting spatiotemporal patterns of apparent commuting travel. At the time this analysis was conducted, discussing it and the various spatial analysis techniques with colleagues at TfL, particularly with those working in operations, was instructive: colleagues at TfL were able to interpret and suggest plausible explanations for many of the different individual-level workplace scenarios that were presented. As a result of these discussions, a further set of visual software was designed for collaboratively exploring the geography of classified workplaces at the scheme-wide level.

Figure 5.8 is an example of this application. Docking stations are again sized according to the number of inbound (in the morning) and outbound (in the evening) commuting

journeys. As in Section 5.3.1, morning (blue) and evening (orange) journeys are distinguished using colour, but this time these journeys are aggregated for all commuting cyclists. Figure 5.8 is effectively a map of 'global workplaces'. At the bottom-right, a slider allows global workplaces to be filtered according to the relative number of morning and evening commutes arriving or departing from these stations. Geodemographic variables appear as vertical bars. The bars change dynamically when data are filtered and can be selected to identify particular subsets of the commuting LCHS population.

The figure shows an interesting observation from this collaborative analysis: that 'global workplaces' with more evening than morning commutes are located entirely towards the periphery of the scheme. This is surprising since these workplaces represent the origin, not the destination, of evening commutes; typically one associates bikes arriving at, rather than departing from, peripheral docking stations during the evening. Exploring this finding with colleagues at TfL, those responsible for LCHS operations drew attention to the fact that the spatial pattern in Figure 5.8 relates very strongly to the initial bounds of the scheme. Typically on weekdays, bikes are disproportionately transported from central London to such peripheral locations during the evening commute. These docking stations often remain full overnight, making it difficult for any commuters wishing to arrive at these peripheral docking stations during the (early) morning commute. As a corollary, during the evening commute docking stations begin to fill up and for those wishing to make outward journeys from these peripheral stations in the (later) evening, bikes are more readily available.

A second finding from exploring the spatial pattern of 'global workplaces' is that the geography of women's workplaces appears to be different from that of men. These spatial differences are best communicated in Figure 5.9. Here, the commuter docking station a member uses the most at peak times is identified and treated as that individual's workplace. London is then divided into $1km^2$ grid cells and the number of 'workplaces' in each cell counted. Differences in the number of 'male' and 'female' workplaces in each cell are compared by calculating signed Pearson's residuals from the Chi-statistic. These residuals are then mapped to a diverging red-blue colour scheme. Red cells contain fewer relative numbers of women's workplaces given the spatial distribution of men's workplaces; blue cells contain greater relative numbers of women's workplaces.

The contrasting geography of men's and women's derived workplaces is perhaps not surprising given the spatial differences identified in the analysis of gender and cycle behaviour (Chapter 4). In that chapter, differences in spatial travel behaviours were

related to a substantial set of pre-existing literature around gendered attitudes to cycling. After discussing the spatial patterns in Figure 5.9 with policy specialists at TfL, it was suggested that these spatial patterns may simply reflect differences in the geography of men's and women's employment centres in London. For example, data modelled from 2011-2012 employment figures show that men fill the majority (65%) of all jobs located in the City of London, whilst in Kensington and Chelsea (south west of Figure 5.9) only 48% of jobs are filled by men (Greater London Authority 2013). This finding is clearly important as it reinforces a factor or control only partially considered so far: that as well as geodemographics and more complex attitudes to cycling, LCHS cyclists' usage characteristics are likely to be a function of *where* those individuals need to travel to access work and other facilities.

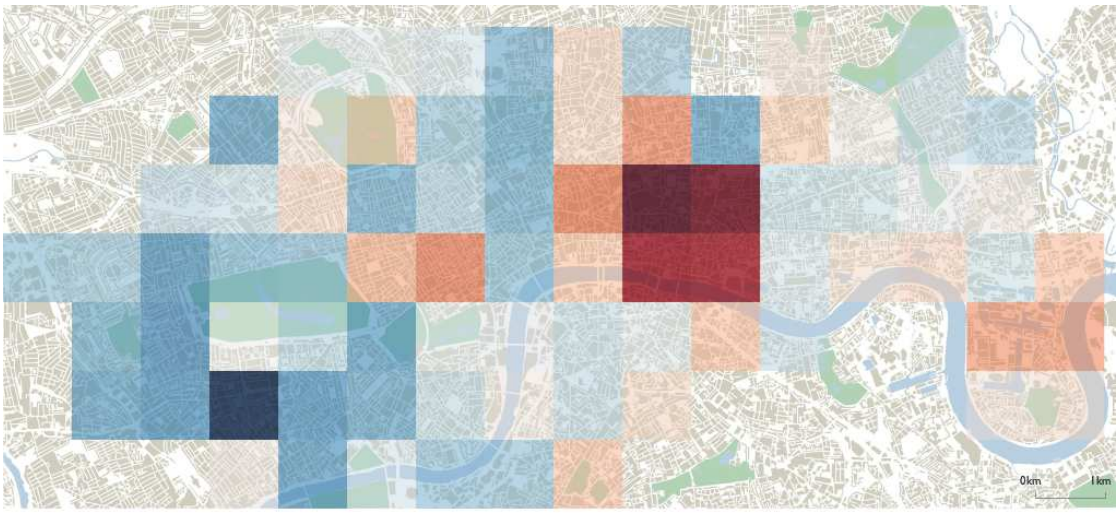

Figure 5.9: The LCHS area is divided into $1km^2$ grid cells. In each cell observed frequencies of 'derived' workplaces for male and female cyclists are recorded and Pearson's residuals from the Chi-statistic are mapped onto a red-blue colour scale assuming equality of proportions between men and women. Grid cells where women's workplaces are over-represented compared to men's are blue; those where women's workplaces are under-represented are red. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

5.5 Discussion

This chapter makes several contributions relevant to the research questions introduced in Chapter 1. Firstly, if chapters 3 and 4 suggest approaches to describing and differentiating LCHS cycle behaviours (RQ1), then the commuter classification is one means through which observed behaviours can be labelled (RQ2). Labelling commuting events with

greater certainty, it is now possible to provide a relatively detailed profile of commuting members and their journeys, as well as investigate emerging themes for analysis, such as interpeak journeys made during members' working day.

Secondly, *visual analytics* techniques and approaches were used to support this more involved analysis activity. With few previous studies examining similar problems using similar datasets, the fact that proposed techniques could be evaluated visually was significant. The visual analytics software created in this chapter enabled deficiencies with suggested techniques to be diagnosed, but also empirically-derived threshold values for the commuter classification to be specified. For example, the problems of spatial outliers displacing workplace centres, of their being more than one workplace centre for many members and of arriving at appropriate threshold values for the density-estimation, were considered through representing individuals' peaktime journey patterns visually.

That the visual encodings used in this software were also relatively intuitive was important; so too was the context under which this software was used. As well as a tool for analysis, the visual analytics systems provided a means through which to communicate to colleagues at TfL any decisions and simplifying assumptions made by the proposed classification techniques. As part of this process, 'domain specialists' at TfL began to openly discuss unexpected spatial patterns and offered new explanations for them. These observations led to a more detailed analysis of the geography of cyclists' derived workplaces and subsequently the observation around men's and women's workplaces. Colleagues at TfL were therefore able to make valuable contributions to very specific aspects of the data analysis.

Finally, RQ3 asks whether identified LCHS cycling behaviours can be explained. In Chapter 4, gendered differences in spatial travel behaviours were related to differing motivations and preferences towards cycling found elsewhere. That these spatial differences exist, even after controlling for how far members live from a docking station, is compelling. However, by identifying members' likely workplaces and then, as a result of interactions with colleagues at TfL, studying the geography of these workplaces, this chapter draws attention to another control that should be considered when attempting to explain customers' spatial travel behaviours: that spatial travel behaviours are also likely to be a function of *where* individuals need to travel to access work or other facilities.

5.6 Moving forward

This chapter represents a progression towards a deeper analysis of behaviours than identified through the exploratory analysis that appears in Chapter 4. Commuting behaviour can now be studied with greater certainty and by labelling all commuting events, new aspects of bikeshare usage, such as interpeak journeys taken as part of cyclists' working day, can also be analysed. Again the techniques used for automatically identifying workplaces were possible because the LCHS dataset provides a total record of behaviour. It would be very difficult to use the same techniques for deriving members' workplaces if only a partial and less spatially and temporally precise set of journey records were available. The next chapter again takes advantage of the completeness of the LCHS dataset when labelling another aspect of behaviour. This time one that has received little attention in Transport Studies, but which may be important for encouraging and initiating cycling uptake: that of group or social cycling.

Chapter 6

Labelling and studying group cycling

Abstract

The group-cycling behaviour of LCHS cyclists is investigated. Group journeys are defined as trips made by two or more cyclists together in space and time. Detailed insights into group-cycling behaviour are generated using specifically designed visual analysis software. In many respects, group-cycling journeys fit an expected pattern of discretionary activity: group journeys are more likely at weekends, late evenings and lunchtimes; they generally take place within more pleasant parts of the city; and between individuals apparently known to each other. A separate set of group activity is found, however, that coincides with commuting peaks and that appears to be imposed onto LCHS users by the scheme's design. Studying the individuals making group journeys, a group of less experienced LCHS cyclists is found that appears to make more spatially extensive journeys than they would do normally while cycling with others; and female cyclists are found to make more late evening journeys than they would do normally when group cycling. For 20% of group cyclists, the first journey ever made through the LCHS is a group journey. This is particularly surprising since just 9% of all group cyclists' journeys are group journeys. Moreover, women are over-represented amongst these 'first time group cyclists'. Turning to the bikeshare cyclists, or bikeshare 'friends', that individuals make 'first time group journeys' with, there is a high incidence of group journeys made with friends of the opposite gender and for a very large proportion (55%) of members, these first ever journeys are made with a friend that shares the same postcode. A substantial

insight, then, is that group cycling appears to be a means through which early LCHS usage is initiated.

Group cycling is a theme that has received very little attention so far within the Transport Studies literature and that would certainly be difficult to investigate using more traditionally-collected datasets. The approach and findings discussed in this chapter represent a new contribution to the Transport Studies discipline; they might be used as a basis for more detailed enquiry.

This work has been published in: Beecham, R. & Wood, J. (2014) Characterising group-cycling journeys using interactive graphics. *Transportation Research Part C: Emerging Technologies*, 47(October), pp.194-206.

6.1 Research context

The ambition in this chapter is to identify and systematically describe group-cycling journeys taken through the LCHS. In this case, group cycling is defined as cycling that happens between two or more individuals together in space and time. A significant contribution is to determine whether certain types of LCHS cyclists are more predisposed to group cycling than others and importantly, whether group journeys are different from the journeys those cyclists typically make. These concerns are reflected in the four research questions below and which guide the analysis covered in Section 6.3:

1. Where are group-cycling journeys, when are they made and who makes them?
2. Are there different types of group-cycling journeys and cyclists?
3. To what extent are group journeys different from the journeys that cyclists typically make?
4. To what extent is group cycling a means through which individuals are introduced to the LCHS?

As yet, there is very little academic research in Transport Studies that has focussed substantively on group cycling. Two small-scale case studies of claimed behaviour that briefly discuss the subject found that respondents reported greater feelings of safety when cycling in groups (Aldred 2012) and that for a small sample of female cyclists, group

cycling was a motivation for returning to cycling (Bonham & Wilson 2012). Whilst it would be possible to study group-cycling behaviour in more detail with larger travel surveys, one reason for the lack of large-scale observational research on the subject may be data availability. Typically in data-driven studies, cycling behaviours are observed by recruiting a small number of self-selected participants and monitoring their travel behaviours over a determinate period of time using GPS (Dill & Gliebe 2008). In order for group-cycling behaviours to be measured using such means, entire social networks would need to be recruited and their behaviours continually monitored, which would likely be problematic.

Despite the lack of research directly measuring group cycling, there is some existing work that is of relevance here. A well-documented barrier to cycling within cities is that of personal safety (Jacobsen 2003). In his 2003 study titled 'Safety in Numbers', Jacobsen (2003) found that collision rates involving walkers and cyclists actually declined as the number of people walking and cycling in an area increased. It is unlikely that those walking or cycling exercise greater caution towards motor vehicles when there are many other walkers or cyclists in an area and Jacobsen (2003) argues that it is motorists' behaviours that are moderated by the increased number of pedestrians or cyclists. It is perhaps reasonable to assume that Jacobsen's (2003) 'Safety in Numbers' thesis also applies to group cycling: group journeys are likely to be more visible than journeys made independently and by extension, group journeys may be materially safer than non-group journeys. That the bicycles available through the LCHS are arguably iconic and conspicuous, and that LCHS users appear as more vulnerable road users – they are less likely than other cyclists to wear helmets or technical clothing (Goodman et al. 2014) – it might be argued that groups of LCHS cyclists moving around London simultaneously may represent a special case of the 'Safety in Numbers' thesis.

Whilst the real safety of cyclists is clearly important, fears about personal safety when cycling actively affect decision-making processes. Importantly, these fears are not experienced evenly across demographic groups, with concerns about cycle safety found to be a greater constraint for women than men (Garrard et al. 2012). Although as the previous chapter suggests, the differing spatial travel behaviours of LCHS men and women may reflect numerous factors, the differences in gendered behaviours described in Chapter 4 are large. The fact that, irrespective of geodemographics, women are routinely under-represented amongst the scheme's most heavy users and appear to preferentially select parts of the city associated with greater levels of safety, may partly reflect the same differing perceptions and attitudes found in many other studies. It might be the

case, as was reported in Aldred (2012), that group cycling positively impacts upon, or helps overcome, the safety concerns of individual LCHS cyclists and subsequently their cycle behaviours. Although it is difficult to make strong inferences about group cycling's impact, it is certainly important to distinguish between the group journeys and non-group journeys of male and female LCHS cyclists, as well as those who are typically less extensive scheme users.

The work discussed in this chapter represents the first large-scale study of group-cycling behaviours of its kind. Again, this new contribution is made possible by the fact that the LCHS provides a total and precise record of behaviour. It would be very difficult within a traditional travel survey to recruit large numbers of complete social networks and monitor the entirety of their travel behaviours. In this respect, the group cycling analysis is perhaps the closest in ambition and substantive content to a 'computational social science' (Lazer et al. 2009) contribution. It is worth remembering, however, that in mining LCHS user records, it is possible only to make *inferences* about group-cycling behaviour. Unlike with survey-based methods, behaviours are not recorded directly. In addition, the analysis covered here misses a particular type of LCHS usage. As discussed in Chapter 3, it is possible to access LCHS bicycles either as a formal member or through paying on the day of travel as a 'casual' user. Casual users generally make around 35% of all LCHS journeys. Analysing journeys made by casual users can be problematic and their group behaviours, or group journeys made between members and casual users, are not examined here. Nor are the group-cycling behaviours of LCHS cyclists and 'normal', non-bikeshare cyclists.

6.2 Data processing

6.2.1 Labelling group-cycling events

The approach to classifying group journeys taken in this chapter is relatively crude. For each member, instances where that person made the same journey (OD pair) with another member at the same time are identified. A two-minute window is used to allow for slight variations in both releasing and docking bikes at a journey's origin and destination. If this 'same journey' happened with a pair of members on more than one occasion, those members are assumed to be bikeshare 'friends' and all same journeys that occur between the two members to be group journeys.

There are problems with this approach. By imposing an exact match on origin and destination, occasions where members cycle with other members for a section of their journey are necessarily excluded. Whilst it might be possible to relax this rule and test only for matching origin station and time, this would lead to greater uncertainty around whether or not journeys are indeed made together and firmer rules may then need to be imposed on the frequency of these events happening between pairs of members. Similarly, a fixed two-minute window may exclude group journeys where, for various reasons, one member takes substantially longer to dock or undock their bike. Clearly the reverse may also be true: that even with relatively strict rules for defining group journeys, two individuals making exactly the same journey on more than one occasion may not be known to one another. It may be possible to derive a model for estimating the probability of these 'non-friend' interactions. However, to be classified as a group journey, a pair of members must make exactly the same journey (within a two-minute window) on more than one occasion. Given these constraints, it might be assumed that those cyclists take the same route and therefore are likely to cycle together.

Running this analysis on the ca. 83,000 members that made journeys in the 12-month study period, just under 20% of members are found to make group journeys and group journeys represent 3% of all 5.09 million journeys. For most members, there is relatively little variation in the size of their group-cycling networks, or the amount of group cycling in which they engage. Ordering group cyclists according to the number of bikeshare 'friends' they have (blue), and group journeys they make (grey), reveals a power-law distribution that commonly exists in social networks (Barabsi & Albert 1999). The majority of cyclists (80%) making group journeys do so with just one other friend and 74% make less than 10 group journeys. Calculating the scale-free exponent (γ) from the two curves using the *maximum likelihood* method, the resulting γ for the number of group journeys made per member is 2.4 and for the number of friends group-cycling members have, is 2.0. A high value for γ, as is found here, suggests a very concentrated network. The probability of finding members with a large number of 'friends' or making a large number of group journeys is small.

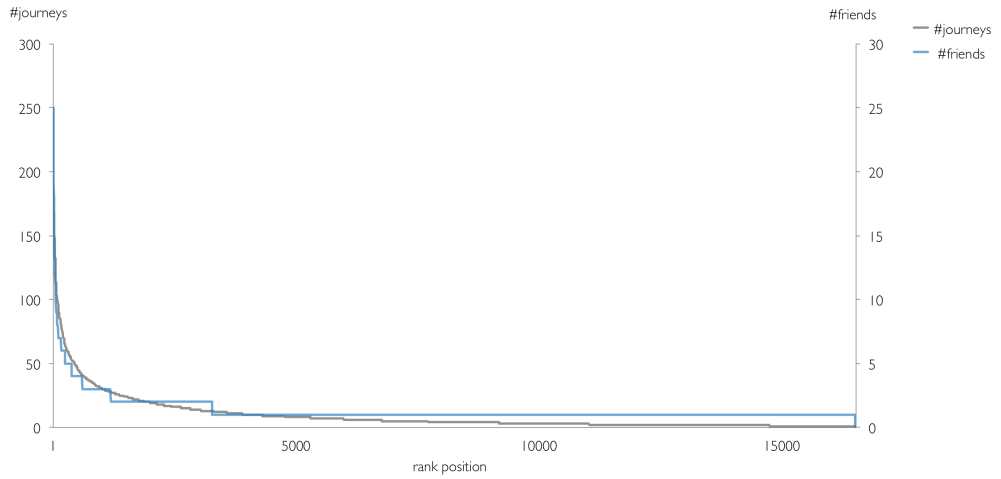

Figure 6.1: Members are ranked according to the number of bikeshare friends they have and group journeys they have made.

Clearly the group-cycling behaviours explored in this chapter are only inferred through analysing LCHS journeys and the possible explanations behind behaviours are informed speculations. One means of validating the group-cycling classification, as well as the commuter classification in the previous chapter, might be to recruit a sample of LCHS cyclists and ask them to recall their usage of the scheme. Aside from the fact that such travel surveys would be difficult to administer, with recall bias perhaps the greatest concern, at present it would not be possible to link surveyed customers with their LCHS usage records for reasons of data privacy.

6.2.2 Visual design

The aim in this chapter is a description of group-cycling behaviours that is both detailed and large-scale. The first two research questions set out above (Section 6.1) are about characterising various types of group-cycling journeys. An important component is also to explore whether particular types of cyclist are predisposed to group cycling, or predisposed to making particular group-cycling journeys. These aims are analogous to earlier work by Slingsby et al. (2013). Here, the authors develop visualization software that enables social communication behaviours to be explored spatially, temporally and by demographic and behavioural category. The design components and interactions that appear in Slingsby et al.'s (2013) work share some similarities with the main exploratory application used in this study. The more general, population-wide group-cycling be-

haviours of LCHS cyclists are therefore explored within the existing software, also used heavily in Chapter 4. Only one addition appears for the group cycling analysis: a 'slider' is added to filter group cyclists according either to the number of group-cycling journeys they make or group-cycling friends they have (Figure 6.3).

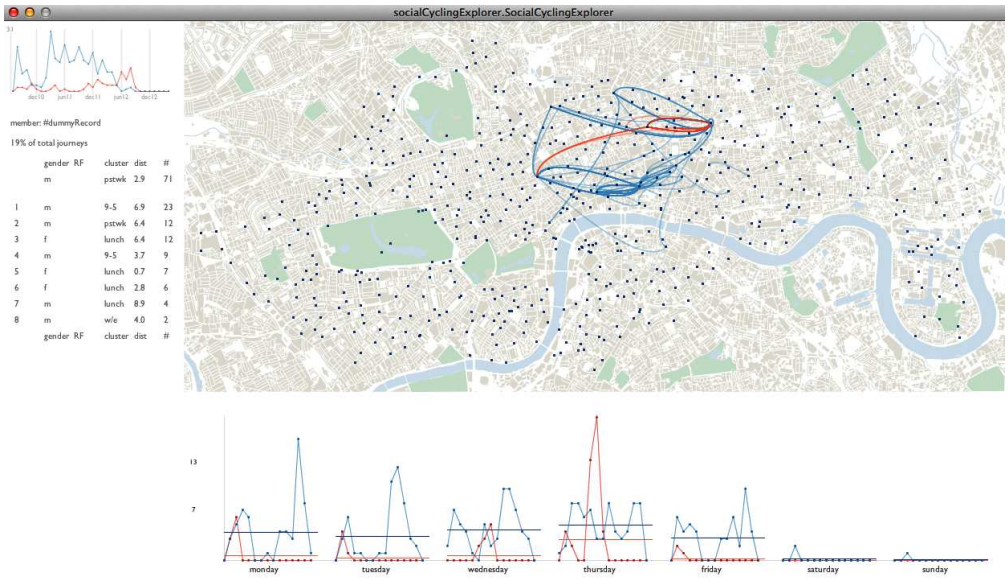

Figure 6.2: A single member is selected and their group and non-group journeys summarised. To preserve customers' anonymity, data for a fictional member and their (bikeshare) social network are generated and displayed here. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

Figure 6.3: Slider for filtering cyclists according to the number of 'friends' they have or group journeys they have made. The light vertical line represents the mid-point. Given the frequency distribution in Figure 6.1, a log scale is used in this slider. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

The strength of the existing tool is that it enables a global overview of cycling behaviours. Slingsby et al.'s (2013) work, however, also allows individual participants' spatiotemporal contact behaviours to be studied. A second application for studying individual members' behaviours and their relationships with other members (Figure 6.2) is developed. The same technique for drawing flow lines, and which was first proposed in Wood et al. (2011), is used in the map view. However, non-group and group journeys are separated using colour hue. In the top left margin, month-on-month volumes of group (red) and non-group (blue) journeys are displayed and below that the selected member and their

bikeshare 'friends' are summarised according to key demographic and behavioural variables. Clicking on any of these friends isolates only group journeys made by these people. In addition, various ordering and filtering techniques are created to facilitate iterative exploration of particular sub-groups of members.

6.3 Analysis

6.3.1 Studying group-cycling behaviour

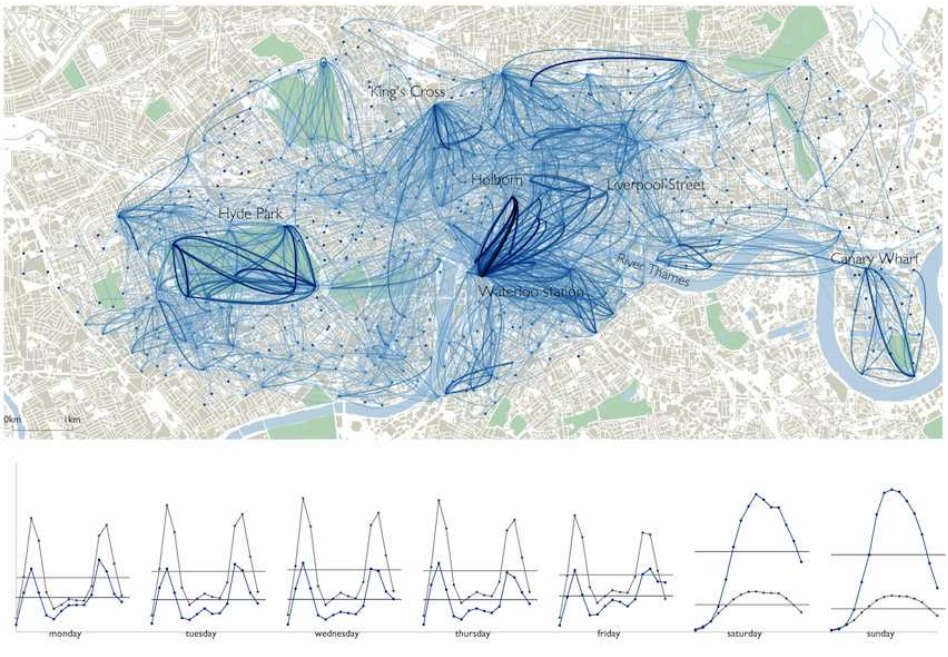

Figure 6.4: Map view: journey lines are weighted according to number of members making journeys. The LCHS's three hub stations, two of which located at major rail terminals (King's Cross and Waterloo), but also in central London (Holborn), are labelled. Temporal view: group journeys appear in blue; all journeys made by the member population are grey. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

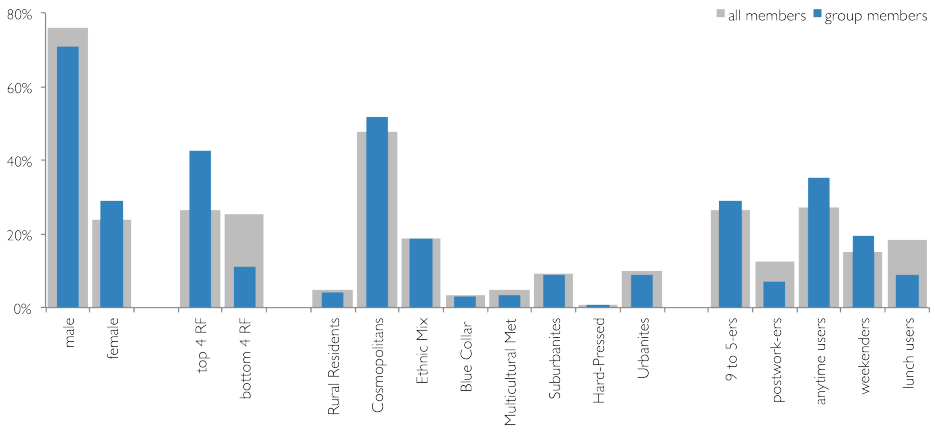

Figure 6.5: Demographic and behavioural profile of group cyclists. From the left: gender, level of usage based on RF groups, 2011 Output Area Classification groupings and temporal cluster membership.

Where are group-cycling journeys, when are they made and who makes them?

Figure 6.4 is a summary of the spatial and temporal structure of group-cycling journeys. Displayed in grey is the relative number of journeys by hour and day of week made in the period 14th September 2011 – 14th September 2012. In blue is the same summary showing only group journeys. Studying the temporal view, it is immediately obvious that the temporal profile of group journeys is very distinct. Compared to total LCHS usage, where 15% of journeys take place at weekends, weekends account for a very large portion (48%) of group-cycling journeys (ϕ_c 0.2, RR 3.1). In addition, a much greater proportion of journeys take place late during Friday and Saturday evenings (between 8pm and 2am): 6% of group journeys happen during Friday and Saturday evenings, whilst this figure for all journeys is 2% (ϕ_c 0.04, RR 2.7). The corollary is that many fewer journeys take place during commuting peaks: 58% for all journeys and 32% for group journeys (ϕ_c 0.1, RR 1.8).

Studying the spatial structure of group journeys, London's parks appear to be a focus of group-cycling activity. A number of journeys can also be found south of the River Thames, within central London and extending east. A very dominant spatial pattern is of journeys between Waterloo rail station and Holborn in central London (labelled in Figure 6.4), where many workplaces rather than shops and other facilities in London are located. Selecting these journeys by performing a spatial filter on the map, a large number coincide with weekday morning peaks. Half of all group-cycling journeys within this region take place between 6am-10am on weekday mornings, whilst this figure for all journeys is just 32%. The reverse is true when only journeys within London's parks are selected: the weekends become especially dominant. Selecting on group journeys that take place within central London itself, peaks are found that coincide with commuting hours, but also lunchtime journeys become particularly prominent towards the end of the working week.

In Figure 6.5, a demographic and behavioural summary of group cyclists is provided. Compared with the total member population, women, high RF scheme users and members living in OAC Cosmopolitan communities, generally affluent inner-city areas, are over-represented amongst group cyclists. Given the spatiotemporal pattern of group-cycling journeys in the previous figure, it is not surprising that the cluster group *weekenders* are over-represented. However, it is also the case that a large portion of group cyclists (29%) are *9-to-5 ers*, a group that typically uses the scheme for commuting purposes.

Are there different types of group-cycling journeys and members?

In some respects, the spatial and temporal structure of group-cycling journeys confirms some preconceptions: the weekends, parks and lunchtimes are a focus of group-cycling activity. However, there is evidence of a variety of group-cycling behaviours. For example, almost a third of all group journeys still coincide with commuting peaks and the main commuter cluster group, *9-to-5 ers*, make up a large portion of group cyclists. Studying these cluster groupings, and the spatiotemporal structure of their journeys, is a useful means of characterising such varying behaviours.

Weekenders are a relatively small group of members (15% of the total member population) typically living within the London area, but who use the LCHS infrequently. Over a quarter of *weekenders* (26%) are group cyclists, greater than for the total member population (20%) (ϕ_c 0.1, RR 1.3). Perhaps as expected, *weekenders*' group-cycling journeys are very concentrated within weekend times (Figure 6.6) and spatially within London's parks and along the River Thames. A noticeable difference between all journeys made by *weekenders* and those journeys that are group journeys, is that group journeys appear more spatially extensive. Significant numbers extend east, as well as along the north and south sides of the river. When selecting non-group journeys made by *weekenders*, however, an extremely dominant pattern is of journeys within Hyde Park. This suggests that, when cycling in groups, *weekenders* generally make a more diverse set of journeys. One way of partially testing this finding quantitatively is to calculate the number of unique journeys, unique OD pair combinations, taken by *weekenders*. Whilst 31% of all journeys taken by *weekenders* are unique, this figure for group journeys is 55% (ϕ_c 0.2, RR 1.8), perhaps suggesting that group journeys for *weekenders* are indeed more diverse than their typical journeys. In terms of demographics, women are very over-represented amongst the *weekenders* who are also group cyclists; they constitute 35% of all *weekenders* but 45% of group-cycling *weekenders* (ϕ_c 0.1, RR 1.3).

9-to-5 ers and *postworkers* together constitute 39% of the total member population. The groups consist of heavy (in the case of *9-to-5 ers*) and occasional (in the case of *postworkers*) commuters often living outside of London. Comparing group journeys with all journeys that these groups typically make, a large proportion coincide with weekends (Figure 6.6). At the same time, however, group journeys for these people do take place during commuting peaks. In fact, group-cycling journeys are even more concentrated within commuting times for *post-workers* than all journeys taken by these members (Figure 6.6). Studying the spatial patterns of these journeys, the hub stations tend to

occupy a large proportion of peak-time group-cycling activity. Journeys either starting or finishing at hub stations comprise 38% of all peak-time group journeys made by *9-to-5 ers*, whereas for all *9-to-5 ers* hub stations only account for 15% of these journeys (ϕ_c 0.1, RR 2.6).

Anytime users are a large group of members who apparently use the scheme for a variety of purposes. Over a third (35%) of *anytime users* are group cyclists. Studying their group journeys, the weekends become particularly prominent, but so too do journeys taken later in the evening on Fridays and Saturdays (Figure 6.6). In fact, *anytime users* comprise 59% of all members making group journeys between 8pm-2am on Fridays and Saturdays. *Anytime* cyclists are generally more active scheme users living relatively close to a LCHS docking station and it perhaps makes sense that those making group journeys later in the evening fit this profile. However, women are over-represented amongst the group cyclists who make these late evening weekend journeys. Whilst they constitute 29% of all group cyclists, 33% of group cyclists making late evening group journeys are women. This is particularly surprising as when considering all journeys (both group and non-group) taking place at these times, women are under-represented: just 22% of members using the scheme in the late evening are women (ϕ_c 0.1, RR 1.5).

Finally, *lunchtime users*, who represent 19% of the member population, consist of generally male, occasional scheme users who live outside of the London area. A relatively small proportion (9%) of these members are group cyclists. Their group journeys coincide with lunchtime hours, but compared with all journeys taken by this group, there is a noticeable peak on Fridays.

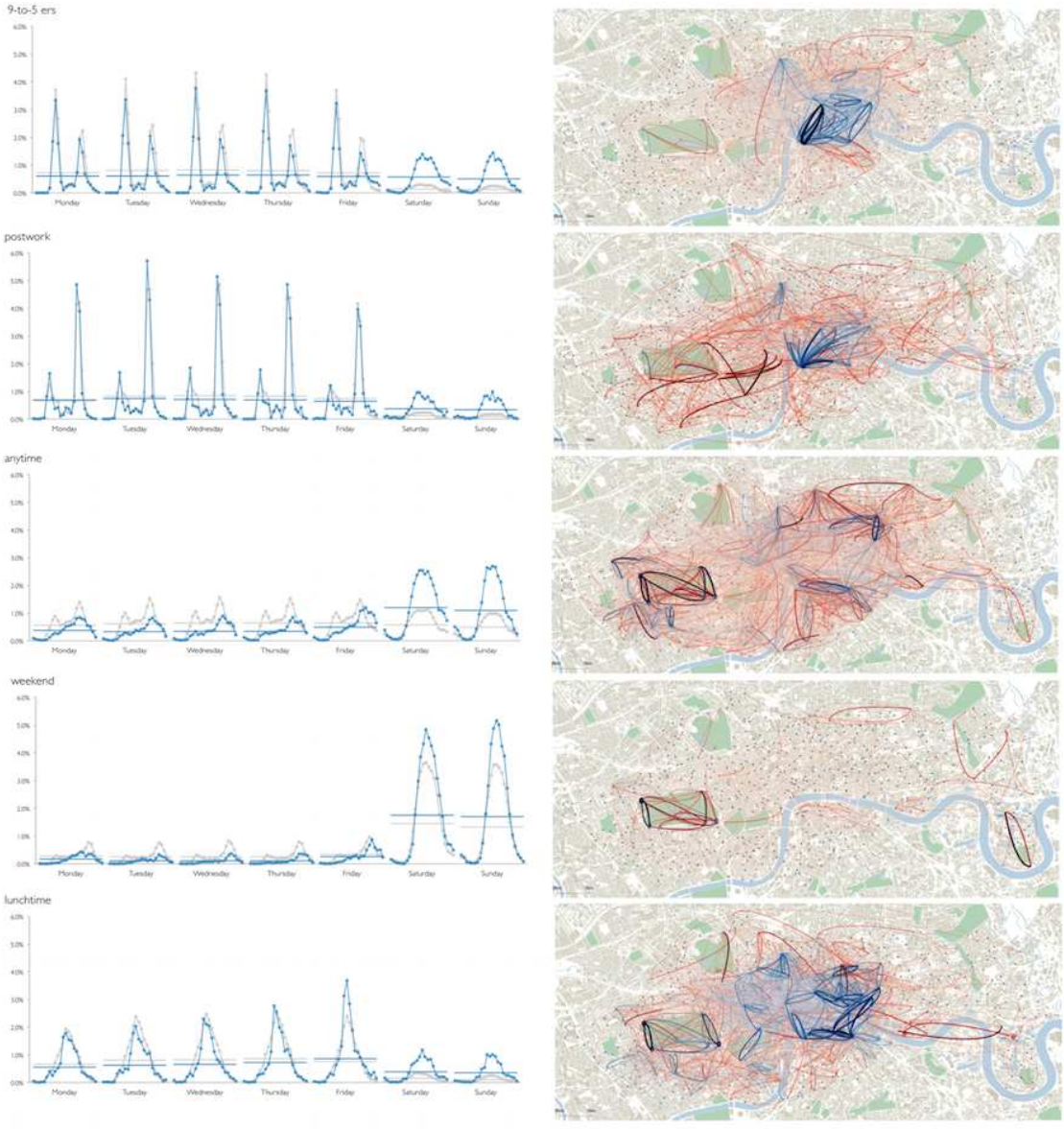

Figure 6.6: Temporal and spatial views of group journeys by cluster membership. Temporal view: all journeys in grey; group journeys in blue. Map view: non-group journeys are blue; group journeys are red. Journey lines are weighted by the number of members making journeys. Since group cyclists represent just 20% of LCHS members, if the same colour scale were used to weight group journeys as non-group journeys it would be very difficult make a comparison; non-group journeys would appear very light and transparent in colour. To enable better comparison, colour weightings for group and non-group journeys are scaled independently. Background mapping uses Ordnance Survey data Crown copyright and database right 2014.

Are group journeys different from the journeys a member typically makes?

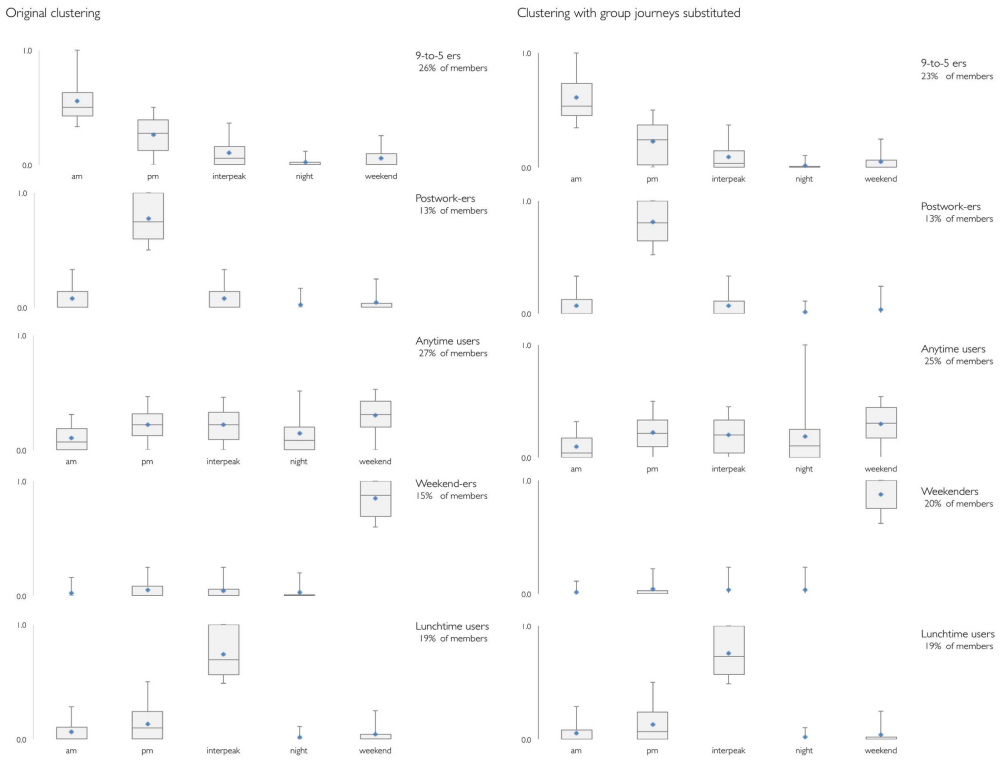

Figure 6.7: Box plots summarising the temporal profile of members in each cluster grouping from the two independent analyses. For each member, the number of journeys made in each time bin is expressed as a percentage of a member's total journeys.

In the previous section, the cluster groupings served as useful descriptors for identifying different types of group cycling. This analysis, and the temporal profiles displayed in Figure 6.6, suggests that differences do exist between the journeys a member typically makes and those that member makes within a group. Visually exploring members' group journeys at an individual level enables a detailed evaluation of these differences. Here, the extent to which group journeys differ at this individual level is studied more systematically by repeating the same clustering procedure described in Chapter 3, but for group cyclists, only their group-cycling journeys are considered.

Although 20% of members are group cyclists, and therefore a number of objects in this second cluster analysis are altered, when run on a 2,000 sample of members (with a representative number of group cyclists), a 5-cluster solution still produces a relatively well-defined set of clusters (ASW = 0.43). Inspecting the temporal profile of these cluster

groupings (Figure 6.7), the same set of labels can be used to describe the output groupings from this 'clustering with group cycling replacement' as in the original cluster analysis. Moreover, of those members who are not group cyclists, only 2% have switched cluster groupings from the original analysis. This suggests that a comparison between these two independent analyses can reasonably be made.

Studying the extent to which group journeys are measurably different, then, 48% of group-cycling members are now in a different cluster grouping than in the original analysis. The largest shift in membership is in the *anytime* user group. Sixty-seven percent of *anytime users* switched cluster grouping, with 51% now classed as *weekenders*. A smaller shift also appears with *9-to-5 ers*, 23% of which are now classed as *weekenders*. These differences make intuitive sense, as it is reasonable to assume that members who use the scheme regularly and for commuting travel would tend to make a larger proportion of their group-cycling journeys at weekends. It is also logical that 14% of *9-to-5 ers* are now classed as *postworkers*. One might expect work-related group journeys to take place during the evening rather than the morning peaks and therefore for a number of *9-to-5 ers* to switch in this direction when filtering only their group journeys. Although this analysis may miss more subtle differences in behaviours, it does provide evidence of the extent to which individuals' group-cycling behaviours are measurably different from non-group behaviours.

To what extent is group cycling a means through which members are introduced to the scheme?

In their study of gender and cycling through the life course, Bonham & Wilson (2012) found that group or social cycling is a means through which adult women are reintroduced to cycling. The same might also be true of LCHS usage: members might be introduced to the scheme first through cycling with others. A database query is therefore performed to find all instances where the first ever journey made by a member is a group journey. The demographic and behavioural characteristics of these 'first-time group cyclists' are then studied to test whether certain individuals are predisposed to this behaviour. In practice, mining the historical journeys dataset one cannot be absolutely certain that the first journey appearing in a member's records is in fact their first ever LCHS journey. As discussed, in addition to registering as a formal member, cyclists can hire bikes as 'casual' users. Tracking returning casual users is problematic and casual journeys that a member may have made before formally registering cannot be easily identified. It should

be noted, then, that when describing the first time a member has used the scheme, only the first journey that individual made *as a bikeshare member* is considered.

In total, this applies to around 20% of all group cyclists. An additional 15% of members made a group-cycling journey within a week of their first ever journey. A significant finding is that women are very over-represented amongst members whose first journey was a group journey. Whilst 29% of group cyclists are women, female members represent 48% of all 'first time group cyclists' (ϕ_c 0.2, RR 1.6). The scheme's least active members, those who use the LCHS rarely and who have not used it recently, are also over-represented amongst members whose first ever journey was a group journey (11% of all group cyclists, but 24% of first time group cyclists, ϕ_c 0.2, RR 2.2).

These findings perhaps suggest that, for a particular type of member, group cycling is indeed a means through which cyclists are introduced to the LCHS. Studying members' bikeshare 'friends' provides further context to these relationships. For each group-cycling member, the individual they cycle with the most – their 'best friend' – is identified, along with all journeys made with this best friend. For 87% of members, the first group journey they made, not their first ever journey as a member, was with this best friend. Since 80% of members have just one bikeshare friend, one would expect this proportion to be large. However, this is particularly true of female group cyclists: 85% of male cyclists' first group journeys were with a best friend, whereas for women, this proportion is 92% (ϕ_c 0.1, RR 1.1). Returning to the idea of 'first time group cyclists', such a difference also exists for those members whose first ever journey is a group journey: 93% of first-time group cyclists' first journeys were taken with their best friend, whilst this was the case for 86% of non first-time group cyclists (ϕ_c 0.1, RR 1.1).

Further analysing these best friend relationships, 34% of male group cyclists' best friends are women and 84% of female group cyclists' best friends are men. The relative probabilities of these two events happening, essentially the female-male split of group cyclists, is 29% and 71% respectively. Female members are therefore more likely to have best friends that are men; and men are more likely to have best friends that are women. These differences again exist when studying first time group cyclists: 89% of female first time group cyclists and 44% of male first time group cyclists made their first ever journey with a cyclist of the opposite gender. An additional, but relevant point is that a large proportion (37%) of group cyclists share exactly the same full postcode as their best friend. Again, this is especially true of first time group cyclists, 55% of which made their first ever journey with a friend sharing the same postcode.

6.4 Discussion

Several insights into the group-cycling behaviour of LCHS members should be highlighted from the analysis discussed in this chapter. Firstly, group journeys do appear to fit an expected pattern of 'leisurely' activity. A large proportion of group journeys take place at weekends, within London's parks and along the banks of the River Thames and analysing the group journeys of *lunchtime users*, Friday lunchtimes are a particular focus of group-cycling activity. In many cases, these group journeys are different from the journeys members typically make and often these differences conform to expectations. For 51% of *anytime users*, 23% of *9-to-5 ers* and 19% of members originally classified as *postworkers*, their group journeys typically fit the profile of the *weekenders* cluster group. Such findings perhaps reinforce the idea that group cycling is discretionary: a planned, leisure-oriented activity taken between members that are likely to be known to each other.

At the same time, however, a different type of group-cycling activity was identified that is perhaps more serendipitous. Studying peaktime group journeys, hub stations were found to be extremely dominant. It is reasonable to question whether, given the prominence of these hub stations and the times at which these journeys are made, 'genuine' planned group behaviour is being measured here. For instance, it could be the case that two individuals do not know each other, but cycle the same route at the same time on more than one occasion merely through chance. Since group journeys are almost identical, having started and ended at the same docking station within a two-minute window, however, one might assume that these two individuals have taken the same route and the benefits conferred by group cycling discussed in Section 6.1 – increased visibility and increased perceived safety from being surrounded by other cyclists – would still apply. These instances, then, perhaps represent a separate category of group activity; one that is partially imposed onto members due to the way the LCHS hub stations are organised. For example, since two of the three hubs are located at major rail stations, situations where a commuter train arrives and causes a surge of competition for bikes at the hub docking station are likely. LCHS operators might then manually replenish this station with bikes, before bikes are immediately withdrawn by a group of waiting members. It is difficult to formally quantify the extent of this 'imposed' group-cycling behaviour. However, in total 11% of group cyclists make group journeys involving hub stations during the weekday peaks, suggesting that this 'imposed' group cycling behaviour may be relatively substantial.

An important motivation for studying group cycling, particularly within an urban context, relates to safety. Given the findings discussed in Chapter 4, where women appear to preferentially select parts of the city that are perhaps associated with greater levels of safety, studying women's group-cycling behaviours was particularly important. It is therefore instructive that women are over-represented amongst group cyclists. In addition, this is true of the *weekenders* cluster group: a collection of members that typically live within the London area, but who are generally inexperienced or infrequent users of the LCHS. There is both visual and quantitative evidence to suggest that *weekenders*' group journeys are more spatially diverse than their non-group journeys and it might be argued that for these people group cycling enables more extensive cycling activity. The same might apply to women who use the scheme late in the evening: women are over-represented amongst late evening group-cycling journeys, but under-represented amongst non-group journeys made at this time.

The possibility of group cycling perhaps enabling scheme usage is addressed directly in Section 6.3.1. For a large proportion of group cyclists (20%), the first journey ever made as a member was indeed a group journey. This is surprising since group journeys constitute just 9% of all journeys made by group cyclists. Women are over-represented amongst 'first time group' members and are also more likely to make first time group journeys with their bikeshare 'best friend' – the person they subsequently cycle with the most. Studying 'best friends' in more detail, there is a very high incidence of best friends of the opposite gender and sharing the same postcode; and this is also true for the friends first time group cyclists make their first journeys with. Especially for women, then, group cycling may help initiate usage of the LCHS and close or immediate friendships may be particularly important to motivating this early scheme usage.

As well as the findings themselves, the group cycling analysis perhaps represents a symbolic contribution. It demonstrates how an aspect of behaviour that would be very difficult to investigate using traditional means can be studied with the LCHS usage data. If group cycling behaviours were to be observed using GPS-based survey methods, for example, entire social networks would need to be recruited and their behaviours monitored continuously over a relatively long period of time, which would clearly be impractical. That the analysis begins to address gaps within current literature, and that findings and implications are discussed in the context of more established theory (Jacobsen's (2003) 'Safety in Numbers' thesis), it might be argued that this chapter is closest in ambition and content to a genuine 'computational social science' study. It makes a novel and substantive contribution to an established social science domain.

6.5 Moving forward

The three analysis chapters (Chapters 4, 5 and 6) and method chapter (Chapter 3) discussed to this point have directly addressed two of this study's research questions. Chapter 4, together with Chapter 3, set out an approach to exploring and describing different LCHS usage behaviours (RQ1) and Chapters 3, 5 and 6 offer approaches to labelling behaviours (RQ2). In studying an aspect of cycling that has so far received little attention, this chapter also makes a contribution to the overall research question: How, and to what extent, can the LCHS dataset be used to contribute to current research in Transport Studies? Each of the findings chapters were nevertheless cautious when offering explanations for the very detailed usage behaviours elicited. Some speculative explanations were given, usually with recourse to existing literature (as in Chapter 4). However, as the spatial analysis of members' workplaces in Chapter 5 suggests, there are clearly a number of factors or controls that might variously affect cyclists' usage behaviours. In the final analysis chapter, an attempt is made to consider individual members' spatial travel behaviours in greater detail. The chapter considers the likely routes cycled between pairs of docking stations and the extent to which the busyness and difficulty of these likely routes might explain differences in individuals' spatial travel behaviours.

Chapter 7

Towards explanation?

Abstract

This final analysis chapter aims to investigate spatial differences in travel behaviour identified in Chapter 4 in more detail. Routing information for every cycled OD pair in the LCHS dataset is estimated using a popular cycle routing engine and from this information, heuristics on the nature of routed journeys are collected. A problem with this approach is that there is no means of knowing how closely estimated routes relate to customers' actually cycled routes. The chapter therefore focusses on an aspect about which there is greater certainty: the bridge that is suggested by the routing algorithm for journeys that involve a river crossing. Differences in male and female cyclists' apparent use of bridges are observed, which appear to relate to the geography of those cyclists' workplaces. Furthermore, studying heuristics for suggested routes over these bridges, there is some evidence to suggest that women may be under-represented amongst commuting journeys that involve a river crossing because those very journeys are associated with relatively busy and demanding routes. These findings are nevertheless quite speculative. Aside from the fact that actual routes are conflated with suggested routes, a number of confounders cannot be easily accounted for: the economic geography of the city, the fact that docking stations are more difficult to use at certain space-times than others and the relative availability of transport alternatives. The chapter reflects on these factors and their implications for *explaining* observed cycling behaviours.

7.1 Research context

In Chapter 4, distinctly different spatial cycling behaviours were found between male and female LCHS users. As well as being less likely than men to use the LCHS regularly and to make apparent utility journeys, women appeared to preferentially select journeys between docking stations located in more 'pleasant' parts of the city – either within parks, or in areas associated with low and slow-traffic streets. These findings are consistent with existing research around men's and women's claimed cycling preferences and one explanation is that they reflect differing attitudes towards cycling held by men and women. This explanation is nevertheless quite speculative. The commuter analysis in Chapter 5 suggests that gendered differences in spatial travel behaviours may also relate to the differing geography of men's and women's workplaces. Another problem is that with only origins and destinations, and no information about the routes taken, it is possible only to guess at the nature of routes that these journeys entailed.

In this chapter, the spatial cycling behaviours of men and women are studied in more detail by collecting estimated routing information for every cycled OD pair in the LCHS dataset. The CycleStreets[1] engine is used to approximate these routes. The ambition is to use this analysis to test the claims made in Chapter 4, but also investigate specific themes of analysis. For example, another finding, only briefly discussed in Chapter 4 is that women are less likely than men to make journeys that involve a river crossing. This finding is true even after controlling for geodemographic and behavioural differences between male and female cyclists. Since London's bridges are generally associated with relatively large, fast-moving roads and with roundabouts or busy junctions at either side, one might hypothesise that such journeys are particularly stressful or demanding; and that this again might explain differences in the relative numbers of men and women choosing to make river crossing journeys. Using detailed information collected from CycleStreets, the chapter aims to address the following questions:

1. Which bridges are most likely to be used by men and women?
2. To what extent are these bridges crossed equally in either direction (northbound and southbound)?
3. Are journeys that involve a river crossing generally more demanding than other journeys made between LCHS docking stations?
4. What are the discriminants of quiet 'estimated' route choice selection?

[1] `www.cyclestreets.net`

This chapter therefore partially attempts to consider LCHS cyclists' route choice. Within Transport Studies, there is a relatively large set of literature that addresses route choice more directly, with both survey-based studies of stated route preference and observational studies of 'revealed' preference. Typically in stated preference studies, respondents are asked to rank different cycle facilities, or suggest a preferred cycle route given a set of pre-defined route options (Bovy & Bradley 1985, Tilahun et al. 2007, Heesch et al. 2012). In revealed preference studies, cyclists might be asked to recall a route that they cycled and those routes compared against a sample of routes generated by a GIS (Larsen & El-Geneidy 2011). Other approaches involve directly observing participants' actually cycled route choices. Menghini et al. (2010) analysed GPS tracks from a comparatively large dataset of 70,000 trips to reconstruct individuals' actually cycled routes. The authors then generated a full set of alternative (non-chosen) routes from which actually cycled routes could be evaluated. Although the volume of data in Menghini et al.'s (2010) study was far greater than in previous studies, data were collected by a private company and not necessarily for the purpose of route preference analysis. No personal information on individuals, for example their gender or age, was revealed and cycle journeys had to be inferred using a mode detection algorithm. A similar approach was taken by Broach et al. (2012) using a smaller GPS dataset of 164 cyclists' journeys, but here the observational data were combined with a richer set of personal attribute information collected as part of a travel survey.

Such structured, experimental studies are clearly highly effective in answering specific questions about perceived (for stated preference) and apparent (for revealed preference) preferences around route choice. Broach et al. (2012), for example, were able to quantify the extent to which distance, turn frequency, slope, the presence or absence of traffic signals and traffic volumes all affect route choice selection. The obvious strength is that in such studies route choices are *known* and therefore only evaluated against a set of directly relevant alternatives. In this study, there is clearly no information on cyclists' *actual* route choices. Only a *likely* route for each cycled OD pair in the LCHS dataset can be generated and the nature of these likely cycled routes studied along with the frequency with which they are made. That the route suggested by the CycleStreets algorithm may be substantially different to the one that is actually cycled is clearly a concern. Whilst one study found that actually cycled routes of commuters rarely deviate from the route suggested by a GIS (Aultman-Hall et al. 1997), another found that on average there is only a 26% overlap between actually cycled routes and GIS routes (Dalton et al. 2013). These latter concerns are one reason for focussing on cyclists' use of bridges.

LCHS cyclists clearly have a limited set of choices for crossing the River Thames and the disparity between cyclists' actual and 'routed' use of bridges may be relatively small.

7.2 Data processing

The CycleStreets engine is used to generate suggested routes for every cycled OD pair in the LCHS dataset. CycleStreets is designed specifically for cyclists and aims to suggest practical routes taking into account the quality of cycling infrastructure, the likely busyness of roads and expected travel times given a number of attributes. The routing algorithm uses various tags collected as part of the OpenStreetMap[1] (OSM) project and working with the Department for Transport, its creators have converted detailed survey data on cycling infrastructure for use in OSM. As a result, the cycle routing considers factors such as: path and surface type and quality, travel time, the presence or absence of signage, the presence or absence of obstacles and traffic calming measures, as well as whether or not a path is lit.

The algorithm works by generating a map of available routes, simplified into a network of straight lines joining nodes (the nodes represent junctions or route start and end points). Distances from the beginning of a journey to all nearby nodes are calculated. For each node, the current distance travelled and route taken is recorded. If it is possible to travel between nodes using a shorter route, then that route is selected as a preferred means of travelling between nodes. The process is repeated until the best route is selected, which is defined as the route that minimises distances between nodes. In order to generate practical routes, various costs are imposed on distances between nodes; and this is how the more detailed attributes collected through OSM are incorporated into the routing algorithm.

CycleStreets allows its users to select one of four route optimisations: shortest path, fastest route, balanced (a mix between travel time and route quietness) and quietest route. For the purpose of this analysis the fastest route is selected. This option may result in bikes being routed on larger or faster roads at certain sections. The reason for selecting this over the balanced or quietest option is that for many, use of the LCHS is occasional and after qualitatively evaluating a small number of routes suggested by CycleStreets, this option appears to suggest more obvious routes that do not require

[1] `www.openstreetmap.org`

extensive familiarity with London's road and cycle network.

CycleStreets provides a web Application Programming Interface (API) to its routing system. Spatial coordinates representing an OD pair are passed to the API through an HTTP request and data on each route returned as XML. Along with a String of coordinates representing waypoints for each route, the following are returned by the API: section length (m), surface type, travel time, turn instructions, count of signalled junctions or crossings and section elevations. Also returned is a 'quietness' score for every planned route. The quietness score ranges from 0% (not at all quiet) to 100% (most quiet) based on a qualitative evaluation of each road or path collected though OSM. Most quiet (quietness score of 100%) are cycle tracks and park paths – generally off-road routes. Slightly less quiet are 'quiet streets', at 75% quietness and shared-use facilities, at 80%. 'Busy roads' are given quietness scores of 50% or less. For each route a single quietness score is provided, taking into account the relative distance travelled on such roads or paths. Routing data are harvested for all ca. 200,000 cycled OD pairs in the LCHS dataset and the processed data stored in an *SQLite* database. To illustrate how the routing data are structured and queried, table schemas appear in Tables 7.1 and 7.2.

oStation	dStation	itinerary	quietness	bridge
154	152	37424975	61	Blackfriars
154	108	37425607	33	Waterloo
372	14	37425611	62	n/a
376	270	37426896	49	Vauxhall
⋮	⋮	⋮	⋮	⋮

Table 7.1: Route summary schema.

itinerary	leg	name	#jncts	#crossgs	turn	coords	elevation	dist	time
37424975	1	Buckley...	0	0		-0.113...	2,2	0,9	9
37424975	2	Mepham...	0	0	lft...	-0.113...	2,2...	0,3...	18
37424975	3	Waterloo...	0	0	rt...	-0.111...	2,2	0,6	2
37424975	4	Alaska St...	0	0	lft...	-0.111...	2,2	0,9...	18
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 7.2: Route section schema.

One of the reasons for collecting these data was that heuristics on estimated routes might be used to make judgements about how challenging particular journeys are. Existing research on safety and cycling infrastructure suggests that the nature and extent of road junctions and roundabouts, the presence of cycle paths, road user speeds and the presence

of right (in the UK and Australasia) and left (elsewhere) turns variously affect both real and perceived levels of safety (Wang & Nihan 2004, Hels & Orozova-Bekkevold 2007). As well as the quietness scores, then, turn instructions and crossings data for each route are processed and stored.

7.2.1 Measurement validity

As discussed, one weakness of the approach taken in this chapter is around the probable mismatch between LCHS users' actually cycled routes and those suggested by the CycleStreets routing engine. One means of partially evaluating how well estimated routes might reflect actually cycled journeys is to make a comparison of travel times. Each routed travel time is evaluated amongst the distribution of actually cycled travel times for each OD pair. This comparison is made on all OD pairs where travel times should tend to a normal distribution: where that journey has been repeated at least 30 times. Each routed travel time for an OD pair (rtt_{od}) is then converted into a $z - score$ given the distribution of actually cycled travel times for that OD pair (att_{od}):

$$z - score = \frac{rtt_{od} - \overline{att_{od}}}{std(att_{od})}$$

One problem with this approach is that travel times should follow a normal distribution. Studying journeys between individual OD pairs and their associated travel times, however, distributions with heavy tails (Figure 7.1) are common. This is not particularly surprising; even amongst bikeshare *members* one would expect discretionary, 'ambling' journeys on many routes. To partially correct for this, for each OD pair of actually cycled travel times, the right tail (95th percentile) is removed before calculating the $z - scores$. Figure 7.1 shows a frequency distribution of these $z-score$ values. Whilst most $z-scores$ lie within a relatively small band of values, the distribution is slightly positively skewed from zero, suggesting that routed travel times are faster than actual travel times, but that this difference is systematic. As well as calculating $z - scores$ for each routed travel time, the skewness and kurtosis for the distribution of actually cycled travel times of every OD pair is calculated. In only 7% of occasions are values for skewness (>2) and kurtosis (>3) very large, suggesting a long right tail. The systematic differences in actual and routed travel times are therefore perhaps not due to the fact that the travel time distributions are non-normal. Instead, the differences might be explained by the fact

that, not included in the routed travel time, is the time spent undocking and wheeling a bike to a road at the start of a journey and returning a bike to its docking station at the end of a journey. In addition, LCHS bikes themselves are very heavy, with a limited number of gears. It is conceivable that the average speeds suggested by CycleStreets for routes are significantly faster than those likely to be cycled using a LCHS bike. For all bikeshare journeys routed using the CycleStreets engine, this average speed is 10 mph, which takes into account factors such as road type, elevation and delays at junctions.

Assuming these factors do overly inflate LCHS travel times, a penalty of 30 seconds is added to the routed travel times (for undocking and docking bikes) and all routed travel times are increased by 10% (to adjust for the weight and nature of LCHS bikes). Doing so has the effect of centring the z-scores (Figure 7.1): 78% of (adjusted) routed travel times lie within one standard deviation of actually cycled travel times for the journeys they aim to represent. If actual travel times were randomly selected from a distribution of travel times for a single OD pair in the LCHS dataset, one would expect 68% of these $z - scores$ to lie within one standard deviation of the mean. This analysis perhaps suggests, then, that routed travel times do relate to the distributions they are supposed to represent and that for a large portion of OD pairs, routed travel times are reasonably close to the centre of these distributions.

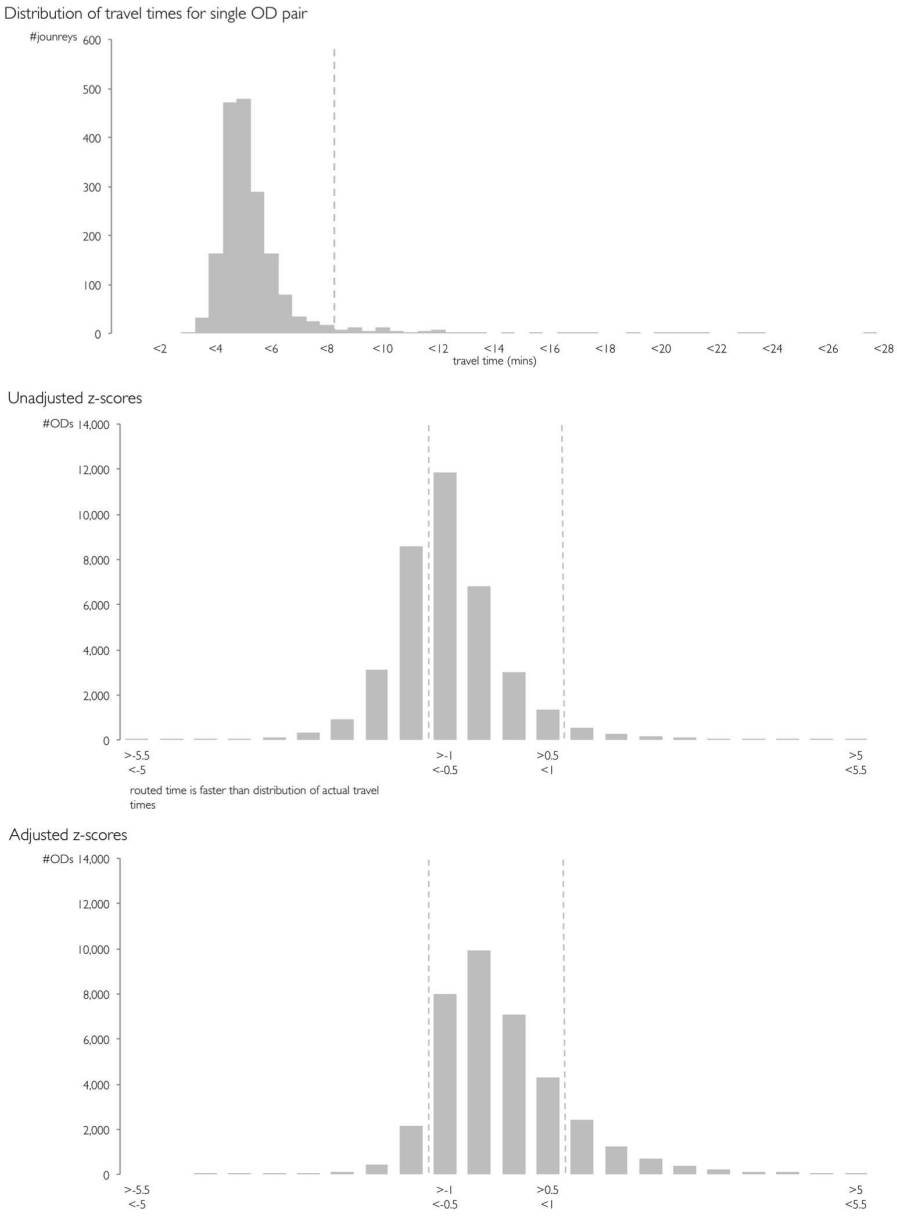

Figure 7.1: Top: Distribution of travel times for a single OD pair. Middle: $z - scores$ calculated for OD pairs where more than 30 journeys are made (and which therefore have a distribution). Bottom: the same data are plotted, but routed travel times are adjusted to account for time spent undocking and docking LCHS bikes and to control for the nature of LCHS bikes.

7.3 Analysis

7.3.1 Suggested use of bridges

Gender and routed use of bridges

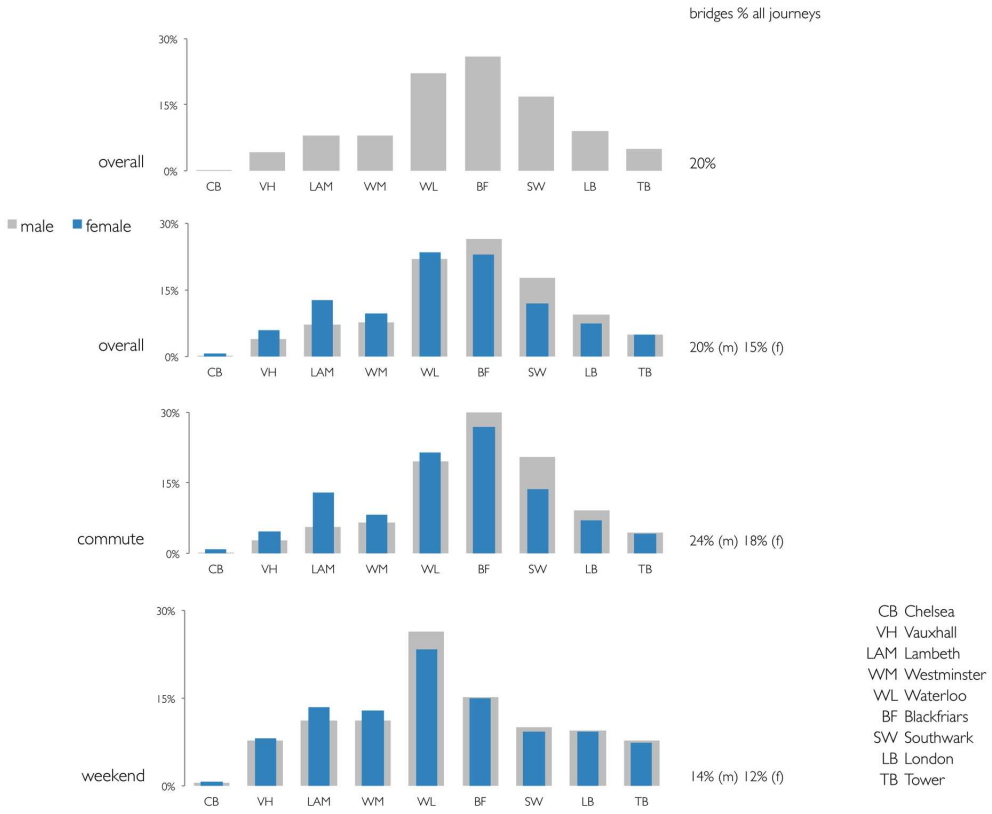

Figure 7.2: Routed journeys over bridges for men and women and by different journey types.

An important finding from the first analysis chapter relates to differences between men's and women's spatial travel behaviours. One insight here is that there are consistently fewer women, and fewer journeys made by women, over bridges than men. In total, 19% of the 5.09 million journeys taken by LCHS members between September 2011 and September 2012 involve a river crossing, with river crossings representing 15% of women's journeys and 20% of journeys taken by men. This is true even after controlling for geodemographic differences in the population of male and female LCHS cyclists. This finding is investigated in more detail here by studying the likely bridges used by LCHS

cyclists, as suggested by CycleStreets.

Commuting and gendered (routed) use of bridges

Figure 7.2 gives relative frequencies for members' estimated use of bridges. In the top row of charts, journeys over each bridge are expressed as a proportion of all journeys involving a river crossing. Below that, the same percentage figures are reported by gender and later by gender and commuter and gender and weekend journeys. The exploratory analysis in Chapter 4 identified large flows around Waterloo, the City of London and Holborn and it is not surprising that, according to the routing algorithm, Waterloo, Southwark and Blackfriars are the most heavily used bridges. Differences between male and female use of these bridges can be easily identified and appear spatially consistent. Relatively more journeys are made by men across bridges close to the City of London and women are over-represented amongst journeys to the west – across Westminster, Lambeth and Vauxhall bridges.

As discussed, there may be a combination of reasons for these differences in the spatial travel behaviours of men and women. An early explanation, made in Chapter 4 with reference to existing literature, might be related to differing preferences and attitudes. A second contributory factor, discussed towards the end of Chapter 5, is that observed spatial travel behaviours must also be motivated by *where* individuals need to travel to access work and other facilities. Returning to Figure 5.9, in which the geography of women's workplaces is contrasted with those of male LCHS cyclists, the gendered differences in men's and women's usage of bridges appear to reflect where LCHS members' jobs are located. This also appears to be true of Figure 7.2. When filtering only by journeys labelled as commutes, the differences between male and female usage of bridges are reinforced. Men are more likely to cross Southwark bridge than women and women are more likely to cross Lambeth bridge than men when *commuting* journeys that involve river crossings are compared. In contrast, there is much greater convergence between men and women when journeys not associated with commuting (weekend journeys) are compared (Figure 7.2).

The geography of LCHS members' workplaces may therefore be a large factor in explaining differences in men's and women's relative usage of bridges. However, it is still the case that women are under-represented amongst all journeys involving a river crossing; and this is especially the case for commuting journeys (Figure 7.2). In addition, although

there is some convergence between men and women at weekends, women still remain slightly over-represented amongst journeys that involve bridge crossings to the west of the city.

Imbalances in directions of travel

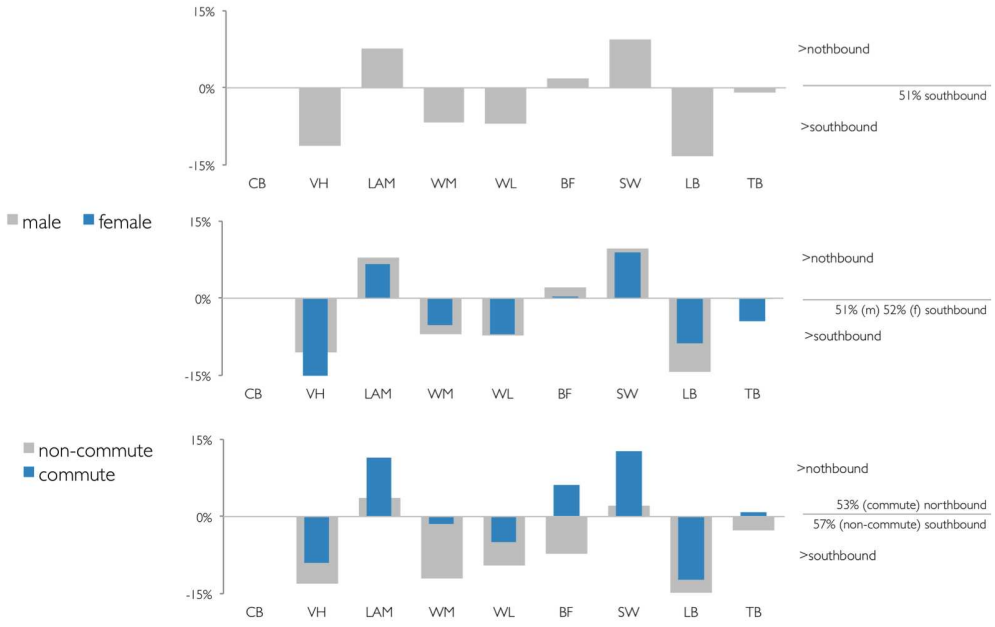

Figure 7.3: The relative balance of southbound-northbound journeys over each bridge, as suggested by CycleStreets.

Certain bridges – Southwark, Blackfriars and Lambeth – tend to be crossed more northbound than southbound and for others the reverse is true. It is rarely the case that there is a perfect balance in the number of northbound and southbound journeys over bridges. There is some convergence between men and women in this respect. Both men and women are more likely to cross Southwark, Blackfriars and Lambeth northbound and are more likely to cross the other bridges southbound. This imbalance is perhaps also related to commuting. Bridges associated with men's commuting (Southwark and Blackfriars) and also women's commuting (Lambeth) are even more likely to be crossed northbound when filtering on commuting journeys and of all commuting journeys that involve a river crossing, 53% of crossings are northbound across the river.

To an extent, this imbalance in favour of northbound commuting journeys might be expected: commuting members make relatively more morning than evening commutes (55% of commuting journeys take place in the morning peak) and workplaces tend to be located north of the river. However, non-commuting journeys that involve a river crossing have an imbalance in the opposite direction: 57% of non-commuting journeys involving a river crossing are southbound journeys.

Routed use of bridges and journey quietness

One explanation for the fact that fewer women make journeys across bridges than men is that the bridges themselves might be perceived to be difficult to negotiate. This is because London's bridges tend to contain relatively fast-moving roads that require riders to negotiate large roundabouts with signalled junctions at either side. This might also be a partial explanation for the fact that there are fewer journeys over bridges at times when more discretionary rather than utility journeys are made – for instance at weekends. Collecting heuristics on the nature of routed journeys, it might be possible to investigate this claim further and identify whether journeys over particular bridges are in fact more demanding than other journeys.

Firstly, frequency-weighted average quietness scores for all journeys that involve bridge crossings are compared with those that do not. Of all actually travelled journeys, those involving a river crossing are in fact associated with slightly higher quietness scores than those that do not, although the difference here is small (52.2 for river crossings; 51.2 for non-river crossings, Cohen's *d.* 0.1). Studying other route heuristics, such as absolute numbers of signalled crossings and right turns, and numbers of crossing and turns per *km* travelled, it appears that journeys involving a bridge crossing are perhaps more technically demanding than other journeys. There is a moderate difference between the average number of signalled junctions or crossings encountered for journeys that involve a bridge crossing and those that do not (4.7 for bridge crossings; 3.3 for non-bridge crossings, Cohen's *d.* 0.7). There is also a small-to-moderate difference between the number of right turns for journeys that involve a river crossing and those that do not (6.4 bridges, 5.3 non-bridges, Cohen's *d* 0.4); although this is not the case when normalising by the distance travelled for these journeys.

There is greater variation in quietness scores when comparing between bridges. Journeys over Southwark, Blackfriars and Waterloo tend to be associated with higher quietness

scores than Westminster, Chelsea, Victoria and Tower Bridge; and journeys over Southwark bridge are associated with particularly high quietness scores (Figure 7.4). Considering the gendered profile of bridge use identified in Figure 7.2, an interesting observation can be made. Journeys over bridges with relatively high levels of female usage, and particularly commuting usage, are in fact less quiet than those associated with men's usage. Assuming these suggested routes in fact do approximate to individuals' actually chosen routes, one might tentatively suggest that women are under-represented amongst river crossing journeys because the routes they must make to commute involve greater levels of risk, or are at least more challenging than those taken by men. This is reflected in the average quietness scores for men's and women's commuting journeys that involve a river crossing: for men this figure is 53.0; for women it is 51.4 (*d.* 0.2).

It is very difficult to provide supporting evidence using the LCHS dataset for formally confirming this claim. The many confounders discussed in Section 7.1 still apply: the geography of members' homes, workplaces or other significant activities; interactions between these locations and the provision and availability of bikes in the LCHS; and also the relative availability of transport alternatives. It is very difficult to account for each of these factors using LCHS dataset alone. However it is the case, when comparing male and female commuting cyclists, that women are less likely to commute across the river than men: 24% of men's commuting journeys involve a river crossing, whereas this value for female members' commuting journeys is 18%. In addition, further support to the claim that suggested route quietness may motivate scheme usage and therefore journey frequency, is the finding that, after excluding commuting journeys, members tend to make more journeys southbound across the river than northbound. This is true even when journeys between the large hub station located at Waterloo are excluded from this analysis. Studying quietness scores for journeys in either direction, these southbound journeys across the river are measurably quieter than northbound journeys (avg quietness = 55.2 southbound; 47.9 northbound, effect size Cohen's *d.* 0.65).

subset	quietness	crossings	crossings/km	rights	rights/km
male	51.4	3.6	1.4	5.5	2.2
female	51.5 (0.02)	3.4 (0.06)	1.3 (0.09)	5.4 (0.06)	2.2 (0.08)
commute	51.7	3.7	1.4	6.0	2.2
non-comute	51.1 (0.06)	3.3 (0.19)	1.3 (0.02)	5.0 (0.37)	2.2 (0.03)
group	52.2	3.1	1.3	5.1	2.2
non-group	51.4 (0.08)	3.5 (0.2)	1.4 (0.13)	5.5 (0.17)	2.2 (0.01)
bridge	52.2	4.7	1.5	6.4	2.0
non-bridge	51.2 (0.10)	3.3 (0.69)	1.3 (0.30)	5.3 (0.44)	2.3 (0.35)
high RF	51.4	3.6	1.3	5.6	2.2
low RF	51.6 (0.02)	3.4 (0.07)	1.4 (0.04)	4.9 (0.28)	2.2 (0.06)
weekend	51.0	3.2	1.3	5.0	2.2
non-weekend	51.5 (0.05)	3.6 (0.16)	1.4 (0.27)	5.6 (0.25)	2.2 (0.06)

Table 7.3: Average quietness scores, number of signalled junctions or crossings and number of right turns for all journeys made by various subsets of the LCHS population. Effect sizes (Cohen's *d.*) between subsets are also reported.

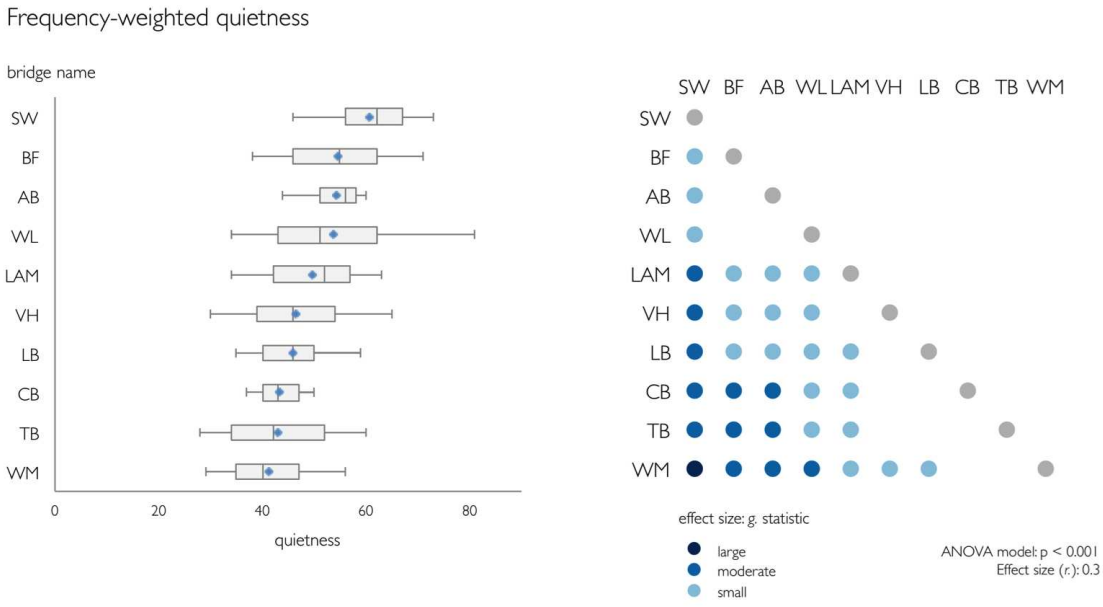

Figure 7.4: Distribution of quietness scores for observed journeys over bridges, as suggested by CycleStreets, are shown as box plots (left) and effect sizes (Cohen's *d.*) for differences between frequency-weighted average quietness scores by bridge appear (right). An Analysis of Variance (ANOVA) model evaluating these differences is also created.

7.3.2 Discriminants of quiet estimated route choice

Relationship between journey frequency and suggested quietness

One hypothesis that might be suggested from the above analysis is that route quietness is positively related to journey frequency: 'quiet' routes are likely to be cycled more frequently than less quiet routes. This is because there are fewer non-commuting, 'discretionary' journeys made northbound across the river, where quietness scores are lower, and fewer commuting journeys made by women that involve a river crossing, also involving lower quietness scores. To test this hypothesis, Pearson's correlation coefficients are calculated on these two variables – quietness and journey frequency – for all OD pairs. Whilst quietness scores are normally distributed, journey frequencies (by OD pair) are very strongly positively skewed. Frequency values are first $log10$ transformed for each OD pair to contrive a more normal distribution. Running correlation coefficients on various geodemographic and behavioural subsets of the member population – on commuting and non-commuting journeys and on group and non-group journeys – there is, however, only a very weak positive correlation (from 0.08-0.18) between journey frequency and quietness score.

That there is so little differentiation in these correlation coefficients, even when filtering on more 'discretionary' journey characteristics such as group cycling, might suggest that individuals' route choice, or rather OD pair choice, is not strongly influenced by route quietness. As discussed, there are various confounders that cannot be easily accommodated within this analysis. Choice or popularity of OD pair is likely to be motivated by that pair of docking stations' visibility or by individuals' knowledge or experience of the scheme; and journeys are likely to be concentrated between parts of the city where particular activities, such work or shopping, take place. With no *a priori* knowledge of individuals' travel requirements or full set of circumstances, and without modelling for the usability of the scheme at particular space-times, it is very difficult to generate an 'expected' model of docking station usage against which observed patterns can be evaluated. There are also of course wider and more fundamental problems of measurement validity – the fact that derived routes are conflated with actual routes. Clearly this enquiry would be more substantial if cyclists' *actual* routes were known; the routing decisions that individual cyclists make could be analysed against a set of non-chosen alternatives, as well as within this framework of wider personal circumstances.

Discriminants of quiet estimated route choice

A final research aim for this analysis was around whether demographic and behavioural variables might be used to predict route quietness. This is partially investigated in Table 7.3. Studying quietness scores alone, however, there is little difference in the journeys made by various behavioural and other groups. Variables such as the number of right turns and river crossings are more discriminating. Journeys involving a river crossing are associated with greater numbers of signalled crossings and right turns (although not when controlling for distance) than those not involving a river crossing. It is also the case that commuting journeys are associated with more signalled crossings and right turns and that the reverse is true of journeys taken at weekends. The effect sizes for these comparisons are nevertheless quite small.

Since the individual heuristics themselves – quietness, turn and crossing frequency – are not particularly discriminating, one means of extending this analysis more formally may be to create a composite measure of route 'stressfulness' that takes into account the three route heuristics appearing in Table 7.3 and use this composite as a dependent variable in a regression analysis. The behavioural and demographic variables appearing in Table 7.3 would then be used as predictor variables. The same confounders discussed in the previous section would nevertheless apply and would need to be accounted for in any proposed model.

7.4 Discussion

This chapter attempted to investigate in greater detail LCHS cyclists' spatial travel behaviours and preferences. Using the detailed routing information collected from CycleStreets, differences in the likely bridges used by male and female cyclists were identified. By studying heuristics for the nature of these journeys, it was possible to add further explanation to the gendered differences identified in Chapter 4. For example, the fact that women are under-represented amongst commuting journeys involving a river crossing may be explained by the possibility that their routed commuting journeys are more busy and challenging than those taken by men. In addition, the imbalance in northbound and southbound journeys during 'discretionary' cycle times might be explained by the fact that routed southbound journeys taken over the river are less challenging than northbound journeys.

As its title suggests, an ambition for this chapter was to progress towards an explanatory data analysis. Whilst the discussed analysis certainly enabled informed explanations to be suggested, the claims again remain quite speculative. An obvious reason for this is that routes are only estimated and not recorded directly. The existing literature around revealed preference introduced at the start of this chapter makes relatively concrete conclusions because participants' route selections – either observed or reported – are known. These known routes can then be evaluated against a set of alternatives to suggest individual preferences. Whilst earlier chapters have discussed deficiencies with more 'traditional' datasets and highlighted opportunities provided by the LCHS dataset, also implicit within them have been the limitations associated with using such passively collected data to analyse social behaviour. A more thorough exposition of these limitations appears in the Conclusion chapter.

An important consideration to emerge out of this, and much of the preceding analysis, is that spatial travel behaviours are likely to be a function of numerous factors: the economic geography of the city, interactions between docking stations at particular space-times, the relative availability of transport alternatives, as well as individual perceptions and attitudes to cycling. As demonstrated in Chapter 4, the size and completeness of the LCHS dataset enables city-level variations in spatial cycle behaviours to be identified. If aspects of individuals' route preferences cannot be investigated without recording routing information directly, it may be possible to take a different approach to studying spatial behaviours and create a model that considers each of the explanatory variables mentioned above in turn and evaluates their effects. For example, it might be possible to generate a model of expected spatial commuting that takes into account, amongst other things, members' home locations, the geography of workplaces in London and the availability of transport alternatives. Observed patterns of commuting in the LCHS dataset, given the classification in Chapter 5, might be compared against this model and the effect of these and extraneous factors quantified. However, like many spatial interaction frameworks (Zhao & Kockelman 2002) this model would necessarily make many assumptions and the existing literature on route preference discussed in Section 7.1 would perhaps still offer the most reliable and substantial research findings.

7.5 Moving forward

This chapter attempted to address RQ3 of this study: To what extent can identified behaviours be explained? Its approach was certainly less exploratory than in previous chapters. The chapter was focussed around a relatively constrained, empirically informed research question: to what extent do the nature of routed journeys explain gendered differences in river crossings? Additional data (estimated routes) were then collected with the aim of answering this research question. Rather than the more speculative, exploratory analysis techniques used in earlier chapters, particularly Chapter 4, a relatively straightforward set of quantitative analysis techniques was also used. In the final chapter, the substantive data analysis that appears through Chapters 3 to 7 is synthesised. The chapter starts by revisiting the three research objectives set out in the Introduction and the extent to which they were achieved. It then articulates this study's main research contributions, which relate both to empirical findings but also the analysis approach. The practical implications for operating bikeshare schemes and policy-related implications for promoting cycle behaviour are then enumerated and the chapter concludes by reflecting on some obvious limitations and opportunities for future research.

Chapter 8

Conclusion

Abstract

This research has met many of its stated objectives. Through detailed spatiotemporal analysis, rich descriptions of cycling behaviour were delineated. That these descriptions relate so strongly to an established set of research suggests that meaningful and perhaps generalisable behaviours can be identified from the LCHS data. The behavioural classifications and substantive analysis in Chapters 5 and 6 is evidence that behaviours can be labelled. This labelling was made possible by the completeness and size of the dataset and many of the classification techniques might not have been repeated in traditional, actively-collected datasets. Throughout this project, hypotheses for explaining behaviours were offered. These hypotheses were made more nuanced by the fact that various spatiotemporal and behavioural controls were investigated using tailored visual analysis software. As the analytical enquiry progressed, and more contextual data or labels were created, the hypotheses and possible explanations became more sophisticated still. There are nevertheless limitations to the LCHS dataset, which make methods for formally quantifying these explanations and hypotheses problematic. This research offers a potential framework and contributes important derived contextual variables for undertaking such formal testing. The research contributions relate both to findings and technique. Chapter 4 provides large-scale evidence to support existing research around gender and urban cycling and the group cycling analysis (Chapter 6) offers several new insights that relate to, and extend, two earlier small-scale studies on the theme. A further academic contribution relates to approach. The LCHS measures use of a new cycle infrastructure and a new user population. As a result, there was some uncertainty around how individual cycling behaviours might be structured and how

they relate to non-bikeshare cycling. By designing flexible visual analysis interfaces, usage behaviours were very quickly explored and characterised. The immediacy of the interactions and the intuitive nature of the designs also enabled colleagues with specialist knowledge of the scheme, but who were new to data analysis, to participate in this research. The same approach might be taken by others working in similar analysis contexts.

8.1 Analysis objectives

This was a data analysis study that aimed to develop a set of research findings that contribute to, and extend, existing research on cycling behaviour within the Transport Studies domain. The analysis sought to identify different cycling behaviours, characterise those behaviours and, considering various spatial, temporal and thematic aspects, suggest motivations behind them. This was reflected in the three research objectives introduced in Chapter 1:

- **Objective 1**: To identify distinct customer cycling behaviours through exploring space-time patterns of travel.
- **Objective 2**: To develop classification techniques for confirming research themes and labelling behaviours identified through exploratory analysis.
- **Objective 3**: To suggest and investigate possible explanations for observed behaviours.

Each of these objectives is now addressed in turn and related to the substantive content of the analysis chapters.

8.1.1 Identifying behaviour

Firstly, the data analysis presented through Chapters 4 to 7 demonstrates that distinct and coherent cycling behaviours do exist and can be characterised. In Chapter 4, differing spatiotemporal cycling behaviours of male and female users were found that are consistent with current research on gender and urban cycling behaviour. The specific findings

and contributions to this literature are enumerated in Section 8.2.1. However, an obvious finding was that women are less likely than men to cycle for commuting purposes and are more likely to preferentially select more 'cycle-friendly' parts of the city. This insight was discovered using the exploratory analysis software and the behavioural variables described in Chapters 3 and 4. Here, customer level data were linked to geodemographic classifiers, a set of behavioural variables for describing customers were precomputed and a visual analysis application was built for exploring the customer-related variables and spatiotemporal aspects of their behaviour. The motivation for building this software was that multiple variables could be compared simultaneously and different spatiotemporal structures of behaviour explored in an immediate and efficient way. An important element of the analysis described in Chapter 4 is that different confounders or controls can be considered. For example, it appears that some of the gendered differences in cycling behaviours relate to differences in the population of male and female customers subscribing to the scheme. Women are over-represented amongst customers who apparently live outside of London and who typically use the scheme heavily and in particular parts of the city. That it was then possible to immediately control for this by filtering according to 'distance from docking station' and also later bikeshare cycling experience, meant that some of the earlier findings could be questioned and more 'fundamental' differences that persist between men and women articulated.

8.1.2 Labelling behaviour

The second objective was to attempt at labelling more formally many of the behaviours identified during exploratory analysis. The work discussed in Chapters 5 and 6 certainly demonstrates that this is possible. The commuter classification in particular allowed specific findings to be investigated with greater certainty. For example, briefly discussed in the exploratory analysis in Chapter 4 were differences in relative numbers of journeys taken interpeak during weekdays. The temporal profile of these interpeak journeys varied when separate spatial selections were made and one hypothesis was that some represented journeys taken during cyclists' working day; others were likely to be 'leisure' trips. By labelling all suspected commuting *journeys* in the dataset, it was possible to then identify interpeak journeys made within cyclists' working day: journeys made by members after commuting into work in the morning or before commuting home from work in the evening.

Each attempt to label behaviour, including the early behavioural variables introduced in Chapter 3, served to augment the LCHS dataset, enabling richer descriptions of observed

behaviour and more detailed hypotheses for explaining behaviour. They also enabled 'gaps' in the customer database that are inevitable in such passively collected data to be partially filled. For example, a consequence of the commuter classification was that a spatial reference for individuals' workplaces was created. Studying the geography of these workplaces in the latter sections of Chapter 5, an important observation emerged: that the geography of LCHS cyclists' derived workplaces differs by gender and appears to relate to differences in the actual geography of men's and women's workplaces in London. This was a useful reminder that spatial differences in cycle journeys may not be essential – they may not reflect wider differences in men's and women's approaches and attitudes to cycling – and that spatial travel behaviours are also likely to be a function of *where* individuals need to travel to access work and other facilities.

The labelling of behaviours enabled new themes of analysis and novel insights to be articulated and this novelty was made possible thanks to the scale, completeness and precision of the LCHS dataset. Chapter 6, on group cycling, is perhaps the most obvious here. It would be difficult to investigate group cycling using traditional, survey-based datasets. If group cycling were to be studied in an observational way, using GPS surveys, entire social networks would need to be recruited, which would be problematic. Whilst group-cycling behaviour is only approximated, and group-cycling journeys made by casual users are absent from this analysis, the group cycling analysis does make a novel contribution to the Transport Studies domain (see Section 8.2.1).

This argument extends to the other classification techniques – the commuter classification and behavioural variables discussed in Chapter 3. Each were again made possible by the completeness and precision of the LCHS dataset. Running the temporal clustering and RF segmentation on a dataset of customers' *claimed*, rather than digitally *observed* or recorded, cycling behaviours would clearly be problematic. Whilst it might be possible for survey respondents to recall the number of bikeshare trips they made over a month-long period, it would clearly be unrealistic to ask about their entire usage of the scheme and more unrealistic still to ask about the specific docking stations they arrived at and departed from, as well as the time in seconds at which these events happened.

8.1.3 Explaining behaviour

Throughout this research, speculative explanations behind observed behaviours were offered and suggested. In Chapter 4, early explanations were made more nuanced by

very quickly investigating different controls and confounders that might also account for differences in men's and women's cycling behaviours. Possible explanations were also validated with recourse to existing literature. As the data analysis progressed, descriptions of behaviours, and therefore possible explanations, became more sophisticated and earlier hypotheses were questioned. Again, most obvious here is that differences in the geography of men's and women's workplaces must also explain differences in their spatial travel behaviours and later that specific aspects of these journeys – the bridges encountered – may also have an effect.

Whilst data-driven explanations were suggested, this study lacks a formal *explanatory data analysis* (Blaikie 2003) chapter and Objective 3 – to suggest and investigate possible explanations for observed behaviours – was only partially met. This was loosely the intention for Chapter 7, which attempted to consider the effect of route difficulty or 'stressfulness' on LCHS cycling behaviours and involved collecting a separate dataset with the aim of answering this research problem. Concerns around measurement validity – around conflating actual with estimated routes – meant that the analysis was relatively modest in ambition. Differences in men's and women's use of bridges and in suggested 'route stressfulness' were outlined and a later section of the chapter attempted to consider the various influences or discriminants of quiet route choice selection more formally. Separate to the problems associated with estimating cycle routes, various confounders that might explain spatial travel behaviours were discussed in this chapter: that choice or popularity of OD pair is likely to be motivated by an individual's knowledge or experience of the scheme, by the usability of the selected OD docking stations and by the fact that journeys will be concentrated between parts of the city where particular activities are located. An expected model of docking station usage controlling for each of these factors, and against which observed behaviours might be evaluated, was suggested. This research makes some progress here, contributing important contextual variables. Developing such a model might be an obvious and immediate challenge for others working with similarly detailed bikeshare data (see Section 8.4.2).

8.2 Research contribution

8.2.1 Thematic contribution

The overriding research question asked: How, and to what extent, can the LCHS dataset be used to contribute to current research on cycling behaviour in Transport Studies? Some of the most substantial domain findings have just been summarised. The tables that appear in this section are used to again list these findings and locate them within the Transport Studies literature to which they aim to contribute. The section is divided in to the two most substantial analysis themes: gendered cycling behaviours and group cycling.

Gendered cycling behaviour

Table 8.1: Contributions of findings to literature on gender and cycling behaviour.

Finding and source	Literature	Contribution
Strong commuting function identified for men and leisure function for women. (Ch. 4).	Survey (Heesch et al. 2012) and observation-based (Dill & Gliebe 2008) studies.	Supporting evidence.
Slower travel times observed for women. (Ch. 4).	Observation-based study (Dill & Gliebe 2008).	Supporting evidence.
Women preferentially select parts of the city associated with slow-traffic streets and cycle lanes offset from major roads. (Ch. 4).	Survey-based (Tilahun et al. 2007, Garrard et al. 2008) and observational (Dill & Gliebe 2008) studies of preferences.	Supporting evidence.
Female cyclists more likely than men to commute in the morning than the evening peaks. (Ch. 4; confirmed Ch. 5).	New insight.	New insight.
Female cyclists less likely than men to make interpeak journeys during their working day. (Ch. 4; confirmed Ch. 5).	New insight.	New insight.

A relatively efficient description of findings and short list of supporting literature appear in Table 8.1. Chapter 4 itself provided detailed descriptions of gendered scheme usage.

These identified behaviours are very consistent with an active set of existing literature, both survey- and observation- based, into gender and urban cycling behaviour. That identified bikeshare cycling behaviours relate so strongly to this literature, and that cycling behaviours are internally consistent, perhaps gives credibility to the new insights that appear in Table 8.1. It is also evidence that distinct and meaningful behaviours do exist and can be identified within the LCHS dataset (Objective 1). At the same time, however, and as discussed in Chapter 7, the findings relating to men's and women's likely route preferences should be argued cautiously. It is also conceivable that some of the 'new insights', for example the fact that women are more likely to commute in the morning rather than evening peaks, will be particular to bikeshare schemes themselves. Their underlying motivations should nevertheless be regarded as relevant to the wider discussion of gender and urban cycling.

Group-cycling behaviour

Table 8.2: Contributions of findings to literature on group cycling.

Finding and source	Literature	Contribution
Especially for less experienced cyclists, group cycling journeys are more extensive in both space and time: for some people, group-cycling may help overcome barriers. (Ch. 6).	New insight related to Aldred's (2012) and Jacobsen's (2003) studies.	New insight/ consistent evidence.
A large portion of group cyclists' first ever journeys were group journeys: group cycling may be a means of initiating cycling. (Ch. 6).	New insight related to Bonham & Wilson's (2012) qualitative study.	New insight/ consistent evidence.
First ever journeys are typically taken with a member of the opposite gender and sharing the same postcode: immediate relationships may be important to initiating bikeshare cycling. (Ch. 5).	New insight related to Bonham & Wilson's (2012) qualitative study.	New insight/ consistent evidence.

The motivation for investigating group cycling came from two studies of cycling attitudes and cultures. Aldred's (2012) qualitative study found that respondents reported greater feelings of safety when cycling in groups and in Bonham & Wilson's (2012) study group cycling was reported as a motivation for returning to cycling having not cycled since childhood. In Chapter 6, a further motivation was set out with reference to Jacobsen's (2003) 'Safety in Numbers' thesis. Since LCHS bikes are relatively conspicuous, it was

argued that groups of LCHS users cycling together in space and time might represent a special case of the 'Safety in Numbers' thesis. The main findings from the group cycling analysis seem highly relevant to Aldred's (2012), Jacobsen's (2003) and Bonham & Wilson's (2012) work: women and less active scheme users appear to make more spatially and temporally varied journeys than they would make normally when cycling in groups; and for a large portion of group cyclists, their first ever journey as a LCHS member was a group journey.

8.2.2 Analytic contribution

Behavioural classifications

Table 8.3: Contributions of techniques to literature on analysing individual travel behaviour.

Technique and source	Literature	Contribution
Recency-Frequency segmentation. (Ch. 3).	Not previously used for analysing traveller behaviour.	Existing technique, new to domain.
Standardised travel time ($z - score$) calculation for individual customers. (Ch. 3).	Similar technique appears in Lathia et al. (2010) and Agard et al. (2011).	Existing technique, new to bikeshare.
Temporal clustering of individual cyclists. (Ch. 3).	Similar technique appears in Lathia et al. (2013) and Agard et al. (2011).	Existing technique, new to bikeshare.
Classifying commuting *events* by deriving individuals' workplaces. (Ch. 5).	Technique and parameters developed are specific to bikeshare context.	Adapted technique, specific to bikeshare context.
Classifying group-cycling behaviour. (Ch. 6).	Technique developed is specific to bikeshare context.	New technique, specific to bikeshare context.

The observation that large, behavioural datasets are relatively new to the Transport Studies domain was made in Section 8.1. There is not a comprehensive literature analysing such timed, OD data, at least for researching individual-level travel behaviours and in Table 8.3 the techniques developed for labelling behaviours are located amongst others within the domain. The Recency-Frequency segmentation introduced in Chapter 3 is a very simple technique that has a long history in database marketing (Novo 2004), but

has to the author's knowledge not before been used in the context of travel behaviour research. The travel time $z - score$ algorithm, which shares similarities with recent work by Lathia et al. (2010) and the temporal clustering of behaviours, again used by Lathia et al. (2013) and Agard et al. (2011), perhaps enable new, more sophisticated 'views' on traveller behaviour (see Section 8.2.2) than in more traditional datasets. Again, to the author's knowledge, they have not previously been used with bikeshare datasets. The commuter classification relies on a commonly used spatial analysis technique (*kernel density estimation*), but was adapted for this research and might be used by others researching commuting behaviour with similar, individual-level origin-destination (OD) data. Finally, the group-cycling classification is an entirely new technique, which might be taken and refined by others.

Applied visualization

A wider analytic contribution relates to the approach taken in this study: to the use of visual analysis software in analysing a large, passively collected behavioural dataset. As discussed in the Introduction chapter, this is one of the first large-scale studies of its kind: one of the first to use the LCHS dataset for researching *individual-level* cycling behaviour. Unlike much of the existing literature, the LCHS usage data were not necessarily recorded for this purpose. There are obvious gaps in the dataset that variously limit the data analysis and potential research questions that might be asked. At the start of the analysis, there was some uncertainty around whether these gaps might be partially filled by leveraging external data or computing derived variables. In addition, as a new dataset recording use of a relatively new cycle facility, it was not certain whether meaningful, individual-level cycling behaviours exist – about the extent to which bikeshare cycling behaviours relate to more general cycling behaviours.

As discussed in Chapter 2 with reference to Sedlmair et al.'s (2012) *design study* paper, visual approaches to analysis are particularly suited to such speculative analysis contexts. For Sedlmair et al. (2012), design studies are applied visualization projects, which start with some data and a domain problem, but where there is usually a degree of uncertainty about the specific aspects of the dataset and research problem that might be studied. Visual analysis techniques are then used to progress the data analysis to a point where both these things are more concrete: where there is a specific set of research questions and the tasks and information used to answer those questions are clear.

Such a trajectory is true of the work described here. By precomputing behavioural variables and designing flexible visual analysis interfaces for exploring their spatiotemporal context, it was possible to very quickly discover numerous usage behaviours. The analysis of gender and LCHS cycling behaviour described in Chapter 4 was conducted entirely within the main set of exploratory analysis software and within a single analysis session. The chapter started with a relatively high-level question – how do male and female cyclists use the LCHS? After very quickly characterising men's and women's behaviours, specific hypotheses and research questions were explored in some detail. The exploratory visual analysis enabled a dialogue with the LCHS dataset and as the analysis chapters progressed, the overall analysis approach became less speculative.

The descriptions of visual analysis process in this study may be relevant to both the Information Visualization and Transport Studies research communities. Whilst the specific visual encodings and methods of interaction were not novel, there are relatively few examples in Information Visualization of their use in such an involved, long-term analysis project (Wood et al. in press) and the description of applied, problem-centred research may be relevant to this community. In terms of the Transport Studies domain, there has been a recent growth in the number of large, public transport datasets recording individual-level behaviours (Bagchi & White 2005, Lathia et al. 2012). The use of visual techniques for exploring and discovering early insights from such data may be of interest to those working with similarly structured datasets and with similarly broad research questions.

A more general case for the described visual analysis approach can be made with reference to current criticisms of analogous, data-driven approaches to social science research that were introduced in Chapters 1 and 2. This work, which typically involves taking secondary, passively-collected datasets measuring human behaviour, has been criticised for being too computer science-focussed: too much emphasis is placed on new and sophisticated computing techniques that are scaleable and not enough on discovering real domain insight (Giles 2012, Watts 2013). One reason is that often such work is led by computer scientists, with little involvement of the social science domain (Giles 2012). Sedlmair et al. (2012) discuss this problem when critiquing *design studies*. For Sedlmair et al. (2012, p. 2436) *'it is essential* [for visualization researchers] *to learn about the target domain and the practices, needs, problems and requirements of domain experts'*. In this study, such a criticism might have been avoided by the fact that the author has some expertise in social science research as well as computing and spent time engaging directly with the transport domain; attending and contributing at conferences and publishing in

transport-related journals. However, an active partner in this research, with specialisms in Transport Operations and Planning, and extremely detailed working knowledge of the scheme itself, was Transport for London (TfL). Although various extraneous factors may explain the success in engaging colleagues at TfL, the visual analysis applications had some effect. Findings were discussed collaboratively with TfL using the main exploratory visual analysis tool. TfL were able to ask specific questions about customers' usage behaviours and these questions were immediately investigated and discussed by interacting with the visual analysis software. As the work on identifying customers' likely workplaces attests (Chapter 5), TfL could contribute even to the more involved analytic activities. Whilst the specific findings of this research may possibly have been reached using non-visual data mining techniques, the creation of interactive visual analysis software perhaps more uniquely supported this context of collaborative analysis and engagement. The documented approach may therefore be used by others working with similarly large, passively collected datasets and who may wish to collaborate across specialist domains.

8.3 Research implications

The academic implications of this research have been discussed. Specific research findings were located within the Transport Studies literature in Tables 8.1, 8.2 and 8.3. That the work has been published in major journals in Transport Studies (*Transportation Planning and Technology* and *Transport Research Part C*), Geography (*Computers, Environment and Urban Systems*) and Information Visualization (*Transactions on Visualization and Computer Graphics*) is further evidence of the domain-specific impact. It was also argued that the visual analysis approach may have wider research implications outside of these domains. A particularly promising domain area might be in health informatics. Patient and clinical data are now increasingly processed and managed digitally. Interactive visual analysis interfaces might be designed to enable healthcare administrators to explore these large data when making resourcing decisions, or clinicians might use such applications to rapidly query and interrogate historical data when considering individual cases (Kamal et al. 2014). Perhaps more closely aligned to the research described in this document, the very large, historical patient data may bring new opportunities for clinical research (Shneiderman et al. 2013, Kamal et al. 2014). Whilst Randomized Control Trials are the 'gold standard' for evaluating clinical interventions, it would be possible to identify from historical data entire cohorts suffering from a particular condition and receiving a particular treatment and at least explore the effects of various controls on patient

outcomes (Shneiderman et al. 2013). Again, the claim of relevance to broader data-driven research might be supported by the fact that invited talks have been delivered at forums separate to Transport Studies and Information Visualization: Health research, Digital Publishing and Journalism (see list of Publications on page vii). Below, two further communities, or rather ambitions, are discussed: attempts within government and elsewhere to promote cycling; and efforts by those working in operations to ensure bikeshare schemes are an efficient and viable transport option.

8.3.1 Promoting cycling behaviour

The large spatial differences in travel behaviours between men and women, and the nature of those differences, suggests that provision of cycling facilities is important for promoting urban cycling amongst women and under-represented cycle groups. Although the possible motivations for these differences were only suggested, that they were observed in such a large population of cyclists is compelling and may help in arguing for investment in cycle facilities or, more generally, in achieving greater gender equality in urban cycling. Evidence of this impact is in written feedback received from colleagues at TfL:

> The analysis of gendered motivations and barriers to using the scheme tell us something about broad cycling behaviour in London. This substantial evidence base will help us secure sustained political and financial support to the ambitious plans set out in the Mayor's 2013 Vision for Cycling in London.
>
> Peter Wright, Senior Cycling Delivery Planning Manager, TfL

A slightly more unique case can be made for incentivising group cycling. There is some evidence to suggest that group cycling may be a means of initiating cycling behaviour and that it might, for certain individuals, support new types of cycling, or journeys that those individuals might not normally make.

Arguments for more specific interventions might be made with reference to the analysis of estimated routes (Chapter 7). The differences in levels of 'quietness' over heavily used bridges identified in Figure 7.4, as well as the frequency with which LCHS cyclists use those bridges at discretionary and non-discretionary times, might suggest that attention should be focussed on improving infrastructure on and around specific bridges: Westminster, Vauxhall and Lambeth. This might be a particular priority if certain types

of behaviour – for example greater levels of utility cycling amongst women – are to be promoted.

8.3.2 Operating bikeshare schemes

A challenge for operators at TfL is to encourage greater levels of usage outside peak times and during the working day. The commuter classification enabled interpeak journeys likely to be made within customers' working day to be labelled. In certain parts of the city, these journeys are concentrated during the lunchtime peaks; in others, they are made more gradually throughout the interpeak period and perhaps by cyclists working at London's universities. The profiles and descriptions of such interpeak activity might be used by those at TfL wishing to incentivise this interpeak behaviour.

One of the most substantial challenges facing bikeshare schemes, and also one of the most cited sources of dissatisfaction with the LCHS (Transport for London 2013), is around bike availability: around being able to collect a bike at a given origin station and easily drop off that bike at an appropriate destination. This is a greater problem at peak times and for those who wish to use the scheme for commuting. Since for each commuting member, a set of docking stations representing their likely 'workplace' is known, it might be possible to send e-mail or SMS alerts suggesting alternative docking stations if members' preferred stations become full at the time they typically commute home.

Finally, by profiling the entire customer population, operators at TfL have a greater understanding of how the scheme is used more generally. The fact that usage data are collected continuously means that different types of behaviours can be monitored as the scheme expands and matures. This information might then be used to inform further expansions or other aspects of the scheme's design (see Section 8.4). Again, below is evidence from TfL of the contribution in this respect:

> Your customer classification and analysis of geographic trends has informed phase 3 of the scheme's expansion into south-west London and intensification of the existing area.
>
> Peter Wright, Senior Cycling Delivery Planning Manager, TfL

8.4 Research limitations and extensions

The limitations of this data analysis study have remained implicit in most analysis chapters. Some of the limitations apply to any research project, some are unique to the LCHS dataset and some relate to the scope and approach taken in this research. Acknowledging what the data analysis does and does not do enables potential gaps to be identified and from there some immediate future research goals can be articulated.

8.4.1 Datasets

The analysis chapters focussed on many of the LCHS dataset's strengths. For example, the fact that the dataset contains a complete, population-level and precise record of behaviour meant that city-wide spatiotemporal cycling behaviours could be characterised and certain themes of analysis examined substantively for the first time. There are nevertheless obvious problems with the LCHS dataset that would not appear in more 'traditional', actively collected datasets.

Demographic detail

Firstly, a lack of demographic information meant that the population structure of LCHS cyclists could not be fully described and compared to a wider cycling population. Whilst respondents' gender is recorded, there is no information on the age and ethnic composition of bikeshare members. It may be the case that bikeshare cyclists are systematically different to 'normal', non-bikeshare cyclists in London or the UK, but it is not possible to quantify this mismatch from the LCHS usage and customer data alone. There is also a separate problem of measurement error here: that, for example, a female bikeshare cyclist may lend their access key to a male cyclist. There is, then, no guarantee that the journeys attributed in the LCHS database to a single member are in fact made only by that member. In addition, since information on individual journey purpose is not recorded directly, the commuting, group cycling and temporal clustering of members were useful shorthands for summarising members and their likely journeys. Clearly, though, such delineations necessarily led to assumptions about cyclists' journey purpose and without direct access to LCHS members there is no obvious way of validating those assumptions.

In future work, these problems might be overcome by actively surveying bikeshare customers. It was not possible in the data-sharing agreement established with TfL to gain access to individual cyclists' contact details. As long-term relationships between researchers and bikeshare data owners develop (Wood et al. in press), however, such information might be made available, or bikeshare operators might make it possible for individuals to volunteer basic information, such as their age and ethnicity, as part of the registration process. A more formal survey that recorded participants' demographic characteristics, but also asked about their typical cycling habits and the types of journeys they make through the LCHS, would also help characterise differences between bikeshare and non-bike cycling. A more intractable problem is that of measurement error. Reflecting on this in their analysis of the LCHS dataset, Goodman & Cheshire (in press) suggest that instances of bikeshare cyclists lending their access keys to others might be rare: the fact that registered members incur substantial charges if a bike is hired for a number of hours and are also liable to pay substantial penalties if a bike is lost or damaged may discourage this sharing of access keys.

Spatial detail

Another deficiency of the LCHS dataset relates to the lack of detailed spatial data. One of the most substantial findings discussed in this document is around the large differences in spatial travel behaviours between men and women. Since only the origins and destinations of bike journeys are known, however, there were obvious limits to the spatial analysis that could be completed; particularly analysis that relates to cyclists' route choice and preferences. Understanding what might motivate these differences, and particularly the extent to which the provision and relative quality of cycle facilities influence behaviours, would be a substantial research task.

As discussed in Chapter 7, a number of observational studies into 'revealed' route preference already exist and on which a study of LCHS trajectories might draw. An obvious challenge for the future will be to investigate whether GPS devices could be attached to at least a sample of LCHS bikes or whether using communities such as *STRAVA*[1], it might be possible to collect volunteered data on actually cycled routes.

[1] `http://labs.strava.com/`

8.4.2 Techniques

The limitations discussed above were generally unavoidable in this project. For example, it was not possible in the data sharing agreement established with TfL to gain access to the full names and addresses of LCHS users and contact them in order to collect their ethnicity, age, occupation and other information. There are, though, some limitations that relate to the nature and scope of the completed analysis.

Study period

Firstly, the research findings are based on a specific 12 months of usage data: from September 2011 - September 2012. The spatial extent of the scheme varied over this 12-month period; the substantial eastern expansion took place in March 2012. Clearly, usage behaviours are partly a function of the provision and availability of bikes and it is reasonable to assume that both LCHS members and their usage behaviours will change as the scheme expands and matures.

In Appendix B, high-level usage behaviours for the period analysed in this research (14th September 2011 – 14th September 2012) are compared with those of the most recent usage data available (27th April 2012 – 27th April 2013). The findings from this initial and cursory analysis remain highly consistent. However, since there is now a reasonable amount of historical data measuring LCHS use, a comparative and longitudinal analysis of behaviours may be a particularly fruitful avenue for further study.

Casual payment cyclists

The analysis undertaken as part of this project misses a substantial aspect of LCHS cycling behaviour – journeys made by casual users. This omission is particularly relevant to the group cycling work: it is conceivable that casual users are particularly predisposed to group cycling and that group cycling journeys taken between casual users and members are also likely.

For every casual user journey in the LCHS journeys database, a numeric variable representing that user's payment card appears. This variable persists over time and it is therefore possible to link the casual journeys of a single user. An initial future analysis activity may be to profile causal cyclists using the same behavioural classifications de-

scribed in Chapter 3. A more substantial, but also difficult task, might be to attempt at linking casual users and members: to identify casual users who then register as formal members. Again, this could only ever be inferred, perhaps by creating a similarity matrix between every casual user and formal member in the dataset. Identifying the point and context under which casual users become formal members may provide important insights to those wishing to increase bikeshare usage as well as promote cycling more generally.

Spatial analysis

Finally, in Chapter 7 some time was spent reflecting on approaches for explaining spatial travel behaviours. Spatial cycling behaviours are likely to be a function of where customers live or depart from within a trip chain, where those individuals then need to travel to in order to access work or other facilities, the relative usability of the scheme at these origin and destination locations and other, more subjective factors, such as attitudes and preferences. Customers' homes are recorded directly in the LCHS dataset, their workplaces and regular origins within a trip chain can be inferred through the commuter classification and an indicator for docking station usability might be derived through temporal analysis of docking station availability (Slingsby et al. 2011). This information might then be used to create a spatial interaction model for expected commuting journeys. Varying the parameters in this model, it might be possible to explore and make quantitative claims about the size of effect contributed by each of these explanatory variables.

8.5 Conclusion

The *thesis* argued in this research is that, as an observational dataset of unprecedented size and spatiotemporal precision, the LCHS offers new opportunities for researching urban cycling behaviour. This thesis is supported by empirical findings, which not only demonstrate that meaningful behaviours exist, but also that new insights can be derived and new contributions made to an established and active set of domain literature. This was a data-driven study that involved working with a large, passively collected dataset that was not necessarily created for the purpose of studying individual-level cycling behaviour. A second argument is that visual analysis approaches are suited to such

speculative research contexts. They enable a detailed space-time context underpinning behaviours to be explored and a dialogue with the dataset that is highly productive. The study's research findings might have been generated without these visual interfaces. However, the immediacy of interactions and intuitive visual encodings were effective in supporting participation from domain specialists with limited analytical expertise, but substantial domain knowledge. In addition to its domain specific contributions, the approach documented in this research might be used by others working in similar analysis contexts.

Bibliography

Agard, B., Morency, C. & Trepanier, M. (2011), 'Mining public transport user behaviour from smart card data', *Science* **333**(6039), 156–157.

Aldred, R. (2010), "On the outside': constructing cycling citizenship', *Social & Cultural Geography* **11**(1), 35–52.

Aldred, R. (2012), Cycling Cultures: summary of key findings and recommendations, Technical report, University of East London, London, UK.

Aldred, R. (2013), 'Incompetent or too competent? Negotiating everyday cycling identities in a motor dominated society', *Mobilities* **8**(2), 252–271.

Anable, J., Schuitema, S., Susilo, Y. & Aditjandra, P. (2010), Walking and cycling in Scotland: Analysis of statistical data and rapid review of the literature, Technical report, NHS Health Scotland, Edinburgh, UK.

Andrienko, N. & Andrienko, G. (2012), 'Visual analytics of movement: An overview of methods, tools and procedures', *Information Visualization* **12**(1), 3–24.

APPC (2013), 'All Party Parliamentary Cycling Group', `allpartycycling.org/`. Accessed: 2013-11-11.

Aultman-Hall, L., Hall, F. & Baetz, B. (1997), 'Analysis of bicycle commuter routes using geographic information systems: Implications for bicycle planning', *Transportation Research Record* **1578**(1), 102–110.

Bagchi, M. & White, P. (2005), 'The potential of public transport smart card data', *Transport Policy* **12**(5), 464–474.

Barabsi, A.-L. & Albert, R. (1999), 'Emergence of scaling in random networks', *Science* **286**(5439), 509–512.

Barker, L. (2009), 'How to get more bicyclists on the road: To boost urban cycling, figure out what women want', *Scientific American* **301**(4), 28–29.

Bartholomew, D. J., Steele, F., Galbraith, J. & Moustaki, I. (2008), *Analysis of Multivariate Social Science Data, Second Edition*, 2 edn, Chapman and Hall/CRC Press, London.

Becker, R. A. & Cleveland, W. S. (1987), 'Brushing scatterplots', *Technometrics* **29**(2), 127–142.

Beecham, R. (in preparation), Using bikeshare datasets to improve urban cycling experience and research urban cycling behaviour, *in* R. Gerike, P. Cox, B. de Geus & J. Parkin, eds, 'The future of cycling', Ashgate, London, UK.

Beecham, R. & Wood, J. (2014*a*), 'Characterising group-cycling journeys using interactive graphics', *Transportation Research Part C: Emerging Technologies* **47**(October), 194–206.

Beecham, R. & Wood, J. (2014*b*), 'Exploring gendered cycling behaviours within a large-scale behavioural data-set', *Transportation Planning and Technology* **37**(1), 83–97.

Beecham, R., Wood, J. & Bowerman, A. (2014), 'Studying commuting behaviours using collaborative visual analytics', *Computers, Environment and Urban Systems* **47**(September), 5–15.

Beroud, B. & Anaya, E. (2012), Private interventions in a public service: An analysis of public bicycle schemes, *in* J. Parkin, ed., 'Cycling and Sustainability', Emerald, Bingley, UK, pp. 269–301.

Bertin, J. (2010), *Semiology of Graphics: Diagrams, Networks, Maps*, ESRI Press, Redlands, California, USA.

Blaikie, N. (2003), *Analyzing Quantitative Data: From Description to Explanation*, 1 edn, SAGE Publications Ltd., London.

Blythe, P. & Bryan, H. (2007), 'Understanding behaviour through smartcard data analysis', *Proceedings of the ICE - Transport* **160**(4), 173–177.

Bonham, J. & Wilson, A. (2012), 'Bicycling and the life course: The start-stop-start experiences of women cycling', *International Journal of Sustainable Transportation* **6**(4), 195–213.

Borgnat, P., Abry, P., Flandrin, P., Robardet, C., Rouquier, J.-B. & Fleury, E. (2011), 'Shared bicycles in a city: A signal processing and data analysis perspective', *Advances in Complex Systems* **14**, 415–439.

Bovy, P. & Bradley, M. (1985), 'Route choice analyzed with stated preference approaches', *Transportation Research Record* **1037**, 11–20.

Broach, J., Dill, J. & Gliebe, J. (2012), 'Where do cyclists ride? A route choice model developed with revealed preference GPS data', *Transportation Research Part A: Policy and Practice* **46**(10), 1730–1740.

Buehler, R. & Pucher, J. (2012), 'Walking and cycling in Western Europe and the United States: Trends, policies, and lessons', *TR News* **280**.

Coe, R. (2002), It's the effect size, stupid: what effect size is and why it is important, *in* 'Annual Conference of the British Educational Research Association (BERA)', University of Exeter, UK.

Cohen, J. (1990), 'Things I have leaned (so far)', *American Psychologist* **45**(12), 11304–11312.

Cohen, J. (1994), 'The earth is round ($p < .05$)', *American Psychologist* **49**(12), 997–1003.

Comacho, T., Foth, M. & Rakotonirainy, A. (2013), 'Pervasive technology and public transport: Opportunities beyond telematics', *IEEE Pervasive Computing* **12**(1), 18–25.

Côme, E. & Oukhellou, L. (in press), 'Model-based count series clustering for bike-sharing system usage mining, a case study with the Vélib' system of Paris', *ACM Transactions on Intelligent Systems and Technology* .

Dalton, A., Jones, A. & Ogilvie, D. (2013), Model behaviour: GPS v GIS to examine our journey to work, *in* 'GIS Research UK (GISRUK) 21st Annual Conference', University of Liverpool, UK.

Davies, D., Gray, S., Gardner, G. & Harland, G. (2001), A quantitative study of the attitudes of individuals to cycling, Technical report, Transport Research Laboratory, Crowthorne, UK.

Davies, D., Halliday, M., Mayes, M. & Pocock, R. (1997), Attitudes to cycling: A qualitative study and conceptual framework, Technical report, Transport Research Laboratory, Crowthorne, UK.

Department for Communities and Local Government (2011), The English Indices of Deprivation 2010: Technical report, Technical report, Department for Communities and Local Government, London, UK.

Department for Transport (2013), National travel survey 2012, Technical report, Department for Transport, London, UK.

Dill, J. (2006), 'Evaluating a new urbanist neighborhood', *Berkeley Planning Journal* **19**(1), 59 – 78.

Dill, J. & Gliebe, J. (2008), 'Understanding and measuring bicycling behavior: A focus on travel time and route choice', *Bicycling* **29**(December), 1 – 70.

Dykes, J. A. (1997), 'Exploring spatial data representation with dynamic graphics', *Computers & Geosciences* **23**(4), 345–370.

Eccles, R., Kapler, T., Harper, R. & Wright, W. (2008), 'Stories in GeoTime', *Information Visualization* **7**(1), 3–17.

Emond, C. R., peaktime, W. & Handy, S. L. (2009), 'Explaining gender difference in bicycling behavior', *Transportation Research Record: Journal of the Transportation Research Board* **2125**(1), 16–25.

Field, A., Miles, J. & Field, Z. (2012), *Discovering Statistics Using R*, SAGE Publications Ltd., London.

Fishman, E., Washington, S. & Haworth, N. (2013), 'Bike share: A synthesis of the literature', *Transport Reviews* **33**(2), 148–165.

Friendly, M. (2009), 'The history of the cluster heat map', *The American Statistician* **63**(2), 179–184.

Froehlich, J., Neumann, J. & Oliver, N. (2008), Measuring the pulse of the city through shared bicycle programs, *in* 'Proceedings of the International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems', Raleigh, North Carolina, USA, pp. 16–20.

Fuller, D., Gauvin, L., Kestens, Y., Daniel, M., Fournier, M., Morency, P. & Drouin, L. (2011), 'Use of a new public bicycle share program in Montréal, Canada', *American Journal of Preventive Medicine* **41**(1), 80–83.

Fuller, D., Sahlqvist, S., Cummins, S. & Ogilvie, D. (2012), 'The impact of public transportation strikes on use of a bicycle share program in London: interrupted time series design', *Preventive Medicine* **54**(1), 74–76.

Garcia-Palomares, J., Gutiérrez, J. & Latorre, M. (2012), 'Optimizing the location of stations in bike-sharing programs: A gis approach', *Applied Geography* **35**(1), 235–446.

Garrard, J., Handy, S. & Dill, J. (2012), Women and cycling, *in* J. Pucher & R. Buehler, eds, 'City Cycling', MIT Press, London, UK, pp. 211–235.

Garrard, J., Rose, G. & Lo, S. K. (2008), 'Promoting transportation cycling for women: the role of bicycle infrastructure', *Preventive Medicine* **46**(1), 55–59.

Gatersleben, B. & Appleton, K. M. (2007), 'Contemplating cycling to work: Attitudes and perceptions in different stages of change', *Transportation Research Part A: Policy and Practice* **41**(4), 302–312.

Giles, J. (2012), 'Computational social science: Making the links', *Nature* **488**(7412), 448–450.

Gleicher, M., Albers, D., Walker, R., Jusufi, I., Hansen, C. D. & Roberts, J. C. (2011), 'Visual comparison for information visualization', *Information Visualization* **10**(4), 289 –309.

González, M. C., Hidalgo, C. A. & Barabási, A.-L. (2008), 'Understanding individual human mobility patterns', *Nature* **453**(7196), 779–782.

Goodman, A. (2013), 'Walking, cycling and driving to work in the English and Welsh 2011 Census: Trends, socio-economic patterning and relevance to travel behaviour in general', *PLoS ONE* **8**(8), e71790.

Goodman, A. & Cheshire, J. (in press), 'Inequalities in the London bicycle sharing system revisited: Impacts of extending the scheme to poorer areas but then doubling prices', *Journal of Transport Geography* .

Goodman, A., Green, J. & Woodcock, J. (2014), 'The role of bicycle sharing systems in normalising the image of cycling: An observational study of London cyclists', *Journal of Transport and Health* **1**, 5–8.

Gordon, G. (2012), Developing methodological approaches to analysing single point bicycle counts, *in* 'Universities Transport Study Group 45th Annual Conference', University of Aberdeen, UK.

Gordon, G. & Parkin, J. (2012), 'Patterns of use by season, day of week and time of day that lead to identifying distinct cycle route typologies', *Cycling Research International* **2**, 104–118.

Greater London Authority (2013), 'Workplace employment by sex and status, borough', `http://data.london.gov.uk/datastore/package/workplace-employment-sex-and-status-borough/`. Accessed: 2014-03-05.

Grolemund, G. & Wickham, H. (submitted), 'A cognitive interpretation of data analysis', *International Journal of Statistics* .

Harrower, M. & Brewer, C. (2003), 'ColorBrewer.org: an online tool for selecting colour schemes for maps', *The Cartographic Journal* **40**(1), 27–37.

Hastie, T., Tibshirani, R. & Friedman, J. (2013), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, 2 edn, Springer, Stanford, California, USA.

Heer, J. & Robertson, G. G. (2007), 'Animated transitions in statistical data graphics', *IEEE Transactions on Visualization and Computer Graphics* **13**(6), 1240–1247.

Heesch, K. C., Sahlqvist, S. & Garrard, J. (2012), 'Gender differences in recreational and transport cycling: A cross-sectional mixed-methods comparison of cycling patterns, motivators, and constraints', *International Journal of Behavioral Nutrition and Physical Activity* **9**(1), 106–118.

Hels, T. & Orozova-Bekkevold, I. (2007), 'The effect of roundabout design features on cyclist accident rate', *Accident Analysis and Prevention* **39**(2), 300–307.

Isenberg, P., Elmqvist, N., Scholtz, J., Cernea, D., Kwan-Liu, M. & Hagen, H. (2011), 'Collaborative visualization: Definition, challenges, and research agenda', *Information Visualization* **10**(4), 310–326.

Jacobsen, P. (2003), 'Safety in numbers: More walkers and bicyclists, safer walking and bicycling', *Injury Prevention* **9**(3), 205–209.

Jensen, P., Rouquier, J.-B., Ovtracht, N. & Robardet, C. (2010), 'Characterizing the speed and paths of shared bicycles in Lyon', *Transportation Research Part D: Transport and Environment* **15**(8), 522 – 524.

Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J. & Banchs, R. (2010), 'Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system', *Pervasive and Mobile Computing* **6**(4), 455–466.

Kamal, N., Wiebe, S., Engbers, J. & Hill, M. (2014), 'Big data and visual analytics in health and medicine: From pipe dream to reality', *Health and Medical Informatics* **5**(5).

Kasik, D. J., Ebert, D., Lebanon, G., Park, H. & Pottenger, W. M. (2009), 'Data transformations and representations for computation and visualization', *Information Visualization* **8**(4), 275–285.

Keim, D. A., Kohlhammer, J., Ellis, G. & Mansmann, F. (2010), *Mastering the Information Age - Solving Problems with Visual Analytics*, Eurographics Association, Goslar, Germany.

King, G. (2011), 'Ensuring the data-rich future of the social sciences', *Science* **331**(6018), 719–721.

Kohavi, R. & Parekh, R. (2004), Visualizing RFM segmentation., *in* 'Proceedings of the 4th SIAM International Conference on Data Mining', Florida, USA.

Kruschke, J. (2013), 'Bayesian estimation supersedes the t–test', *Journal of Experimental Psychology: General* **142**(2), 573–603.

Kusakabe, T., Iryo, T. & Asakura, Y. (2010), 'Estimation method for railway passengers' train choice behavior with smart card transaction data', *Transportation* **37**(5), 731–749.

Larsen, J. & El-Geneidy, A. (2011), 'A travel behavior analysis of urban cycling facilities in Montréal, Canada', *Transportation Research Part D: Transport and Environment* **16**(2), 172–177.

Lathia, N., Ahmed, S. & Capra, L. (2012), 'Measuring the impact of opening the London shared bicycle scheme to casual users', *Transportation Research Part C: Emerging Technologies* **22**, 88–102.

Lathia, N. & Capra, L. (2011), How smart is your smartcard? Measuring transport behaviours, perceptions and incentives, *in* '13th ACM International Conference on Ubiquitous Computing', Beijing, China.

Lathia, N., Froehlich, J. & Capra, L. (2010), Mining public transport usage for personalised intelligent transport systems, *in* 'IEEE 10th International Conference on Data Mining (ICDM)', Sydney, Australia.

Lathia, N., Smith, C., Froehlich, J. & Capra, L. (2013), 'Individuals among commuters: Building personalised transport information services from fare collection systems', *Pervasive and Mobile Computing* **9**, 643–664.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. & Alstyne, M. V. (2009), 'Computational social science', *Science* **323**(5915), 721–723.

Menghini, G., Carrasco, N., Schüssler, N. & Axhausen, K. (2010), 'Route choice of cyclists in Zurich', *Transportation Research Part A: Policy and Practice* **44**(9), 754–765.

Miller, H. J. (2010), 'The data avalanche is here. Shouldn't we be digging?', *Journal of Regional Science* **50**(1), 181–201.

Munzner, T. (2008), Process and pitfalls in writing information visualization research papers, *in* A. Kerren, J. Stasko, J.-D. Fekete & C. North, eds, 'Information Visualization', Vol. 4950 of *Lecture Notes in Computer Science*, Springer, Berlin, Germany, pp. 134–153.

Niemeier, D. A. (1996), 'Longitudinal analysis of bicycle count variability: Results and modeling implications', *Journal of Transportation Engineering* **122**(3), 200–206.

Novo, J. (2004), *Drilling Down: Turning Customer Data into Profits with a Spreadsheet - Third Edition*, 3 edn, Jim Novo, St. Petersburg, FL, USA.

Ogilvie, F. & Goodman, A. (2012), 'Inequalities in usage of a public bicycle sharing scheme: Socio-demographic predictors of uptake and usage of the London (UK) cycle hire scheme', *Preventive Medicine* **55**(1), 40–45.

O'Sullivan, D. & Unwin, D. (2002), *Geographic Information Analysis*, John Wiley & Sons, New Jersey, USA.

O'Brien, O., Cheshire, J. & Batty, M. (2014), 'Mining bicycle sharing data for generating insights into sustainable transport systems', *Journal of Transport Geography* **34**(January), 262–273.

Parkin, J., ed. (2012), *Cycling and Sustainability*, Emerald, Bingley, UK.

Pelletier, M., Trépanier, M. & Morency, C. (2011), 'Smart card data use in public transit: A literature review', *Transportation Research Part C: Emerging Technologies* **19**(4), 557–568.

Perer, A. & Schneiderman, B. (2008), Integrating statistics and visualization: Case studies of gaining clarity during exploratory data analysis, *in* 'Proceedings of the 26th annual SIGCHI conference on Human factors in computing systems', New York, USA.

Phan, D., Xiao, L., Yeh, R., Hanrahan, P. & Winograd, T. (2005), Flow map layout, *in* 'Proceedings of the IEEE Symposium on Information Visualization', Minneapolis, Minnesota, USA.

Pooley, C., Tight, M., Horton, D., Scheldeman, G., Jopson, A., Mullen, C. & Chrisholm, A. (2011), Understanding walking and cycling: Summary of key findings and recommendations, Technical report, Lancaster University Environment Centre, Lancaster, UK.

Priolli, P. & Card, S. (2005), The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis, *in* 'International Conference on Ingelligence Analysis', McLean, Virgina, USA.

Pucher, J. & Buehler, R. (2012), *City Cycling*, MIT Press, London.

Páez, A., Trépanier, M. & Morency, C. (2011), 'Geodemographic analysis and the identification of potential business partnerships enabled by transit smart cards', *Transportation Research Part A: Policy and Practice* **45**(7), 640–652.

Rae, A. (2009), 'From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census', *Computers, Environment and Urban Systems* **33**(3), 161–178.

Reed, W. J. (2001), 'The Pareto, Zipf and other power laws', *Economics Letters* **74**(1), 15–19.

Robbins, N. B. (2005), *Creating More Effective Graphs*, Wiley-Blackwell, New Jersey, USA.

Roberts, J. C. (2005), Exploratory visualization with multiple linked views, *in* J. Dykes, A. M. MacEachren & M.-J. Kraak, eds, 'Exploring Geovisualization', Elsevier, Oxford, UK, pp. 159–180.

Robinson, A. (2008), Collaborative synthesis of visual analytic results, *in* 'IEEE Symposium on Visual Analytics Science and Technology, 2008', Columbus, Ohio, USA.

Roethlisberger, F., Dickson, W. & Wright, H. (1967), *Management and the worker: an account of a research program conducted by the Western Electric Company, Hawthorne works, Chicago*, Harvard University Press, Cambridge, Massachusetts, USA.

Rousseeuw, P. J. (1987), 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics* **20**, 53–65.

Rugg, G. & Petre, M. (2007), *A gentle guide to research methods*, Open University Press, Berkshire, UK.

Sedlmair, M., Meyer, M. & Munzner, T. (2012), 'Design study methodology: Reflections from the trenches and the stacks', *IEEE Transactions on Visualization and Computer Graphics* **18**, 2431–2440.

Shaheen, S. A., Guzman, S. & Zhang, H. (2012), Bikesharing across the globe, *in* J. Pucher & R. Buehler, eds, 'City Cycling', MIT Press, London, UK, pp. 183–210.

Shekhar, S., Evans, M., Kang, J. & Mohan, P. (2011), 'Identifying patterns in spatial information: A survey of methods', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**(3), 193–214.

Shen, Z. & Ma, K.-L. (2008), MobiVis: a visualization system for exploring mobile data, *in* 'IEEE PacificVIS 2008', Kyoto, Japan, pp. 175–182.

Shneiderman, B. (2002), 'Inventing discovery tools: Combining information visualization with data mining', *Information Visualization* **1**(1), 5–12.

Shneiderman, B., Plaisant, C. & Hesse, B. W. (2013), 'Improving healthcare with interactive visualization', *Computer* **46**(5), 58–66.

Slingsby, A., Beecham, R. & Wood, J. (2013), 'Visual analysis of social networks in space and time using smartphone logs', *Pervasive and Mobile Computing* **9**(6), 848–864.

Slingsby, A., Wood, J. & Dykes, J. (2011), Visualizing bicycle hire model distributions, *in* 'Geoviz Hamburg', Hamburg, Germany.

Sun, L., Axhausen, K., Lee, D.-H. & Huang, X. (2013), 'Understanding metropolitan patterns of daily encounters', *Proceedings of the National Academy of Sciences* **110**(34), 13774–13779.

Thomas, J. & Cook, K. (2006), 'A visual analytics agenda', *IEEE Computer Graphics and Applications* **26**(1), 10–13.

Thomas, T., Jaarsma, R. & Bas, T. (2013), 'Exploring temporal fluctuations of daily cycling demand on Dutch cycle paths: The influence of weather on cycling', *Transportation* **40**(1), 1–22.

Tilahun, N. Y., Levinson, D. M. & Krizek, K. J. (2007), 'Trails, lanes, or traffic: Valuing bicycle facilities with an adaptive stated preference survey', *Transportation Research Part A: Policy and Practice* **41**(4), 287–301.

Tin, S. T., Woodward, A., Robinson, E. & Ameratunga, S. (2012), 'Temporal, seasonal and weather effects on cycle volume: an ecological study', *Environmental Health* **11**(1), 12.

Transport for London (2013), 'Barclays Cycle Hire customer satisfaction and usage survey: Members, wave 7 (Q3 2013/2014)', `http://www.tfl.gov.uk/corporate/publications-and-reports/cycling-and-walking`. Accessed: 2014-06-05.

Tufte, E. R. (1986), *The visual display of quantitative information*, Graphics Press, Cheshire, Connecticut, USA.

Tukey, J. (1962), 'The future of data analysis', *The Annals of Mathematical Statistics* **33**(1), 1–67.

Tukey, J. (1980), 'We need both exploratory and confirmatory', *The American Statistician* **34**(1), 23–35.

Tukey, J. W. (1977), *Exploratory Data Analysis*, 1 edn, Addison-Wesley, London.

Tukey, J. & Wilk, M. (1966), Data analysis and statistics, an expository overview, *in* 'International Workshop on Managing Requirements Knowledge', Los Alamitos, California, USA.

Vickers, D. & Rees, P. (2006), 'Introducing the area classification of output areas', *Population trends* (125), 15–29.

Vogel, P., Greiser, T. & Mattfeld, D. (2011), 'Understanding bike-sharing systems using data mining: Exploring activity patterns', *Procedia `[U+FFFD]`ocial and Behavioral Sciences* **20**, 514–523.

Wagenmakers, E.-J. (2007), 'A practical solution to the pervasive problems of p–values', *Psychonomic Bulletin and Review* **14**(5), 779–804.

Wang, Y. & Nihan, N. (2004), 'Estimating the risk of collisions between bicycles and motor vehicles at signalized intersections', *Accident Analysis and Prevention* **36**(3), 313–321.

Watts, D. (2013), 'Computational social science: Exciting progress and future directions', *The Bridge on Frontiers of Engineering* **43**(4), 5–10.

Wickham, H., Cook, D., Hofmann, H. & Buja, A. (2010), 'Graphical inference for infovis', *IEEE Transactions on Visualization and Computer Graphics* **16**(6), 973–979.

Wood, J., Beecham, R. & Dykes, J. (in press), 'Moving beyond sequential design: Reflections on a rich multi-channel approach to data visualization', *IEEE Transactions on Visualization and Computer Graphics* pp. 1–10.

Wood, J., Radburn, R. & Dykes, J. (2010), vizLib: Using the seven stages of visualization to explore population trends and processes in Local Authority research, *in* 'GIS Research UK 18th Annual Conference (GISRUK 2010)', University College London, UK.

Wood, J., Slingsby, A. & Dykes, J. (2010), 'Visualisation of origins, destinations and flows with OD maps', *The Cartographic Journal* **47**(2), 117–129.

Wood, J., Slingsby, A. & Dykes, J. (2011), 'Visualizing the dynamics of London's bicycle hire scheme', *Cartographica* **46**(4), 239 – 251.

Woodcock, J., Tainio, M., Cheshire, J., O'Brien, O. & Goodman, A. (2014), 'Health effects of the London bicycle sharing system: Health impact modelling study', *BMJ: British Medical Journal* **348**.

Yang, T., Haixiao, P. & Qing, S. (2011), Bike-sharing systems in Beijing, Shanghai, and Hangzhou and their impact on travel behavior, *in* 'TRB 90th Annual Meeting', Washington DC, USA.

Yuill, R. (2011), 'The standard deviational ellipse: An updated tool for spatial description', *Geografiska Annaler. Series B, Human Geography* **53**(1), 28–39.

Zhao, Y. & Kockelman, K. (2002), 'The propagation of uncertainty through travel demand models', *Annals of Regional Science* **36**(1), 145–163.

Appendix A

Technical Notes

Data analysis was carried out using freely available and open source software: the Processing development environment (`http://www.processing.org`), the statistical programming language R (`http://www.r-project.org`) and the serverless database software library SQLite (`http://www.sqlite.org`). Various software libraries, particularly the giCentre Utilities library (`http://www.gicentre.net/utils/`), were used to support, amongst other things, drawing statistical graphics and zooming and panning in the visual analysis applications. Software were developed using the Eclipse IDE (`http://www.eclipse.org`) and statistical analysis and data mining procedures using RStudio (`http://www.rstudio.com`). This document was written in LaTeX, using the TeXShop (`http://pages.uoregon.edu/koch/texshop/`) editor.

Appendix B

Comparison with April 2012 - April 2013 dataset

There is much consistency between waves. However, members are perhaps more active in the April 2012-2013 wave. More journeys are made per head; the relative size of the commuting population has increased by 4% points; the relative size of the more active cluster groups -- *anytime* users and *9-to-5 ers* -- has increased by 1% point; and men now constitue 77% of the user population. There is very little difference in the temporal structure of journeys -- in hourly flows by day of week -- save for the fact that there are very slighly fewer made in the Friday evening peak. This perhaps makes sense given the fact there are slightly fewer (-1% point) *postworkers* in the 2012-2013 wave. Although the spatial structure of journeys is consistent, flows in the eastern expansion area are more prominent and the large Waterloo-Holborn flows are less prominent. That there are very slightly more members (+1% point) living in the most deprived IMD quintiles and very close to a docking station also suggests greater number of members from within the eastern expansion area (see also Goodman & Cheshire, in press).